UNIVERSITY OF BELGRADE
FACULTY OF PHILOSOPHY

Vanja N. Subotić

LINGUISTIC COMPETENCE AND NEW EMPIRICISM IN PHILOSOPHY AND SCIENCE

Doctoral Dissertation

Belgrade, 2023

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЗОФСКИ ФАКУЛТЕТ

Вања Н. Суботић

ЈЕЗИЧКА КОМПЕТЕНЦИЈА И НОВИ ЕМПИРИЗАМ У ФИЛОЗОФИЈИ И НАУЦИ

Докторска дисертација

Београд, 2023

Information about the Supervisor and Dissertation Defense Committee

Supervisor

Dr Miljana Milojević, Associate Professor
Department of Philosophy
University of Belgrade—Faculty of Philosophy

Dissertation Defense Committee

Dr Ljiljana Radenović, Full Professor
Department of Philosophy
University of Belgrade—Faculty of Philosophy

Dr Voin Milevski, Associate Professor
Department of Philosophy
University of Belgrade—Faculty of Philosophy

Dr Vladan Devedžić, Full Professor
Department of Software Engineering
University of Belgrade—Faculty of Organizational Sciences

Dr Luca Malatesti, Full Professor
Department of Philosophy
University of Rijeka—Faculty of Philosophy

Date of the dissertation defense: _____

ACKNOWLEDGMENTS

I did not really feel like a grown-up when I turned 18. I did not feel like a grown-up when I got my first job, nor when I moved out. In all these instances I did not have the impression that I have changed in any particular way or became, as they say, "my own man" (or woman, for that matter). Now I do.

Teeny-tiny parts of the mosaic that you have in front of you have been presented at numerous conferences. I was lucky to receive valuable feedback at *LICPOS* in 2023, *Trust in Science Summer School* in 2022, *ESHS Early Career Workshop*, *EENPS Biannual Conference*, and *Speaking Bodies Conference* in 2021, *EECP Annual Workshop for Early Career Scholars* in 2020, *SILFS Postgraduate Conference* and *Formal Methods in Science and Philosophy* in 2019. While being a demonstrator for the undergraduate course *Philosophy of Mind* at the Faculty of Philosophy for the past four years, I presented some of the arguments to my students, and got the opportunity to learn from their questions (that were more of comments). At times when I felt a bit under the weather, their still fresh zeal reminded me of my love for philosophy—and I cannot thank them enough for that.

The time has come to drop some names, though. I have been blessed with two families— each being challenging in their peculiar way—one at home and one at work. And this is *not* a cheesy hustle culture reference. First, I am indebted to my parents, Marica & Nikola, and my older sister Ivana (AKA third parent), for their immense love, kindness, compassion, and support. During the last five years, while I was struggling to produce *one* doctoral dissertation, my sister basically effortlessly brought into this world *two* amazing children, Vasilija & Konstantin, so I contend that she remains the most successful person in the Subotić household. At least, until Vasilija and Konstantin come of age. Second, I am grateful to Janko Nešić, Petar Nurkić, Željko Mančić, Jelena Pavličić, Ljiljana Radenović, Slobodan Perović, and Nada Dumić for their encouragement and genuine friendship (and all our pub quiz escapades and hard-earned 13th places). I was also pretty lucky to have my partner Branko Pantić beside me during the final stages of writing.

Finally, the moment you were all waiting for – a more official address and tone. I thank my supervisor, Miljana Milojević, for her patience (and, boy, was she patient!) and for having confidence that we will make something out of nothing, given the scarce philosophical literature on deep learning when I started doctoral program in 2019. Interestingly enough, the first person I met from the Department of Philosophy was her. Back in 2013 when I was in high school and attended an event for the future philosophy students in Students' Cultural Centre, I remember seeing a twentysomething on the podium, with a long braid, wearing canary yellow jacket, and being thrilled that I'd be learning from someone looking so fierce. The thrill has not faded during these 10 long years of acquaintance. Her comments, as well as the comments of the committee for my *viva* of dissertation proposal in 2021—namely, Vladan Devedžić, Voin Milevski, Miloš Vuletić, and Vojislav Božičković—were of significant help for setting me straight, and any remaining mistakes or weak points are my responsibility only.

ABSTRACT

The topic of this dissertation is the nature of linguistic competence, the capacity to understand and produce sentences of natural language. I defend the empiricist account of linguistic competence embedded in the connectionist cognitive science. This strand of cognitive science has been opposed to the traditional symbolic cognitive science, coupled with transformational-generative grammar, which was committed to nativism due to the view that human cognition, including language capacity, should be construed in terms of symbolic representations and hardwired rules. Similarly, linguistic competence in this framework was regarded as being innate, rule-governed, domain-specific and fundamentally different from performance, i.e., idiosyncrasies and factors governing linguistic behavior. I analyze state-of-the-art connectionist, deep learning models of natural language processing, most notably large language models, to see what they can tell us about linguistic competence. Deep learning is a statistical technique for the classification of patterns through which artificial intelligence researchers train artificial neural networks containing multiple layers that crunch a gargantuan amount of textual and/or visual data. I argue that these models suggest that linguistic competence should be construed as stochastic, pattern-based, and stemming from domain-general mechanisms. Moreover, I distinguish syntactic from semantic competence, and I show for each the ramifications of the endorsement of connectionist research program as opposed to the traditional symbolic cognitive science and transformational-generative grammar. I provide a unifying front, consisting of usage-based theories, construction grammar approach, and embodied approach to cognition to show that the more multimodal and diverse models are in terms of architectural features and training data, the stronger the case is for the connectionist linguistic competence. I also propose to discard the competence vs. performance distinction as theoretically inferior so that a novel and an integrative account of linguistic competence originating in connectionism and empiricism that I propose and defend in the dissertation could be put forward in scientific and philosophical literature.

Keywords: *Linguistic Competence*, *Natural Language Processing*, *Connectionism*, *Empiricism*, *Nativism*, *Deep Learning*, *Large Language Models*.

Scientific field: *Philosophy*.

Scientific subfield(s): *Philosophy of Language, Philosophy of Mind*, *Philosophy of Science*.

Сажетак

Тема ове дисертације је језичка компетенција—способност разумевања и продуковања израза на било ком природном језику. Браним емпиристичку позицију у погледу језичке компетенције, уско повезану са конекционистичком когнитивном науком. Ова линија когнитивне науке је била супротстављена традиционалној симболичкој когнитивној науци упареној са трансформационо-генеративном граматиком, која се обавезала на нативизам у погледу порекла когнитивних процеса. Разлог за обавезивање на нативизам лежао је у томе што је људска когниција, укључујући језичку способност, у овој парадигми схватана као вођена симболичким репрезентацијама и унапред одређеним правилима. У том духу, језичка компетенција је представљена као урођена, доменоспецифична и фундаментално различита од језичког понашања. Кроз анализу савремених конекционситичких модела за процесирање природног језика базираних на дубоком учењу, или, још специфичније, великих језичких модела, испитујем њихову корисност за разумевање природе језичке компетенције. Дубоко учење је статистичка техника разврставања и препознавања шаблона на основу великог броја текстуалних и/или визуелних података путем којих се обучава вишеслојна вештачка неуронска мрежа имплементирана у конекционистички модел. Аргументујем да ови модели показују да је боље лингвистичку компетенцију схватити као стохастичку, вођену шаблонима, и доменогенералну. Уз то, разликујем синтаксичку од семантичке компетенције, и за сваку експлицирам последице које потичу од прихватања конекционистичке насупрот симболичке когнитивне науке. Стварајући уједињени фронт од конекционизма, теорија базираних на употреби, конструкционе граматике и утеловљених приступа когницији, који би резултовао у мултимодалним моделима завидног диверзитета у погледу података и архитектуралних детаља унутар самих модела, показујем да се конекционистичка језичка компетенција може бранити као успешнија од симболичке за симулирање људске језичке способности. Најзад, предлажем да се одбаци генеративистичка дистинкција између језичке компетенције и понашања, као теоријски инфериорна у односу на ново, интегрисано виђење језичке компетенције проистекло из конекционизма и емпиризма које представљам у дисертацији.

Кључне речи: *Језичка компетенција*, *процесирање природног језика, конекционизам*, *емпиризам*, *нативизам*, *дубоко учење*, *велики језички модели*.

Научна област: *Филозофија.*

Ужа научна област: *Филозофија језика, филозофија духа, филозофија науке.*

CONTENTS

List of Figures, Tables, and Images

List of Abbreviations: Terminology

AI — Artificial intelligence

ANN — Artificial neural network

CNN — Convolutional neural network

CTM — Computational theory of mind

CAP — Construction grammar approach

DCNN —Deep convolutional neural network

DL — Deep learning

FNN — Feedforward neural network

GOFAI — Good-Old-Fashioned Artificial Intelligence

LLM — Large language model

LOT — Language of thought hypothesis

LSTM — Long short-term memory

ML — Machine learning

NLP — Natural language processing

PDP — Parallel distributed processing

RNN — Recurrent neural network

SRN — Simple recurrent neural network

TGG —Transformative-generative grammar

UBT — Usage-based theories

XAI — Explainable artificial intelligence

List of Abbreviations: Primary Sources

Nota bene: Bibliographical details of primary scholarly sources can be found at the end of the dissertation within **References**.

CSM — John Cottingham, Robert Stoothoff, and Dugald Murdoch's translation and edition of *The Philosophical Writings of Descartes*

Essay — Peter H. Nidditch's revision and edition of Locke's *Essay concerning Human Understanding*

New Ess. — Peter Remnant and Johnatan Bennett's translation and edition of Leibniz's *New Essays on Human Understanding*

T — L. A. Selby-Biggess's edition and Peter H. Nidditch's revision of Hume's *Treatise on Human Nature*

EHU — L. A. Selby-Biggess's edition and Peter H. Nidditch's revision of Hume's *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*

הִנְנִי הָעָנִי מִמַּעַשׂ

Hinneni he'ani mima'as.

Here I stand, impoverished in my deeds and merit.

0. INTRODUCTION

Hate me, hate my dog.

— Jerry Fodor (1990: xii)

You are reading the sentence on this piece of paper and understand, in the course of reading, the meaning of all the syntactic constituents of this sentence—its subject, predicate, etc. This is probably because you have a good command of English or English is your mother tongue. Mine is Serbian. You and I both have linguistic competence, the capacity to produce and understand sentences in our mother tongue or any foreign language we have learned. This also means that our mind is currently processing language, often called *natural language* so that it can be distinguished from formal languages in logic and mathematics or programming languages for coding. What is the nature of this special power we have? What cognitive mechanisms underlie such power? How does it operate? What exactly constitutes it? These are the questions that I will be examining within this dissertation.

The issue of whether one should look inside or outside our heads for unraveling the nature of our cognitive capacities, including linguistic competence, is an old one, or, better yet, an old-fashioned one. To tackle it, one has to take the road many philosophers—as well as few cognitive scientists, linguists, and AI researchers—have travelled on. Historically, in the early modern period, the two camps of philosophers, namely rationalists and empiricists, proposed two quite different images of our nature. Rationalists (Descartes 1628/1988, 1641/1988, 1644/1928, Leibniz 1704/1981) insisted on the innate ideas that the Lord himself bestowed upon us. Language, specifically, was considered as the innate gift *par excellence*—how else could we be set apart from animals were it not for our language faculty that allows for codifying moral principles and religious dogma? Empiricists (Locke 1690/1975, Hume 1740/1978, 1748/1975), however, held that most of our knowledge begins with the senses since we learn to navigate the world by having more or less direct sensory contact with it. Language, specifically, was understood as a set of perceivable signs standing for the content in our heads that societies bestowed upon us. The later cohorts of philosophers sympathetic to rationalism (Frege 1892/1952, Montague 1970a/1974, 1970b/1974) strived toward the perfect language that would avoid the pitfalls of the natural language full of ambiguities. Those endorsing the aspects of empiricism (Putnam 1975, Burge 1979) preferred natural language with all its ambiguities because it provided us with an unflattery but authentic mirror of our nature.

A couple of centuries later, the rationalist dream was realized within the philosophy of language in the analytic tradition with the advent of the new formal language of propositional and predicate logic. Thus, formal language should have served to replace natural language and to give a more precise and rational account of word and sentence meaning that did not go beyond the cognizing mind of the individual. The other camp of philosophers of language wanted to go beyond the cognizing mind of the individual, into the environment and community, to locate the meaning of words and sentences. The early

modern philosophers were concerned with origins of knowledge, including knowledge about grammar of our own language, while philosophers of language tried to locate the very process of naming things or referring to things—it is either in associating descriptions to objects of reference (Russell 1908), or in causal chains linking the object of reference to the one who baptized it first (Kripke 1972). Then came scientists with all their methodology and implicit or explicit philosophical inclinations.

In the 1970s, cognitive science, a novel interdisciplinary field that was constituted by philosophy, linguistics, biology, Artificial Intelligence (AI), anthropology, and psychology, was inaugurated. Thus, inherited theoretical commitments of its constituents became embedded in the methodology of cognitive science conceived as multi-disciplinary field (Miller 2003). Moreover, such commitments entailed what frameworks, tools, and explanations would be deemed acceptable for investigating linguistic competence. Traditional symbolic cognitive science preferred formal language of mathematical logic as a means to express and understand human cognitive processes, used the tools of the Good-Old Fashioned Artificial Intelligence (GOFAI) to model such processes through symbolic representations and discrete rules for manipulating such representations, and, consequently, considered only deductive-nomological explanations as true explanations of cognitive phenomena (see Fig. 1).

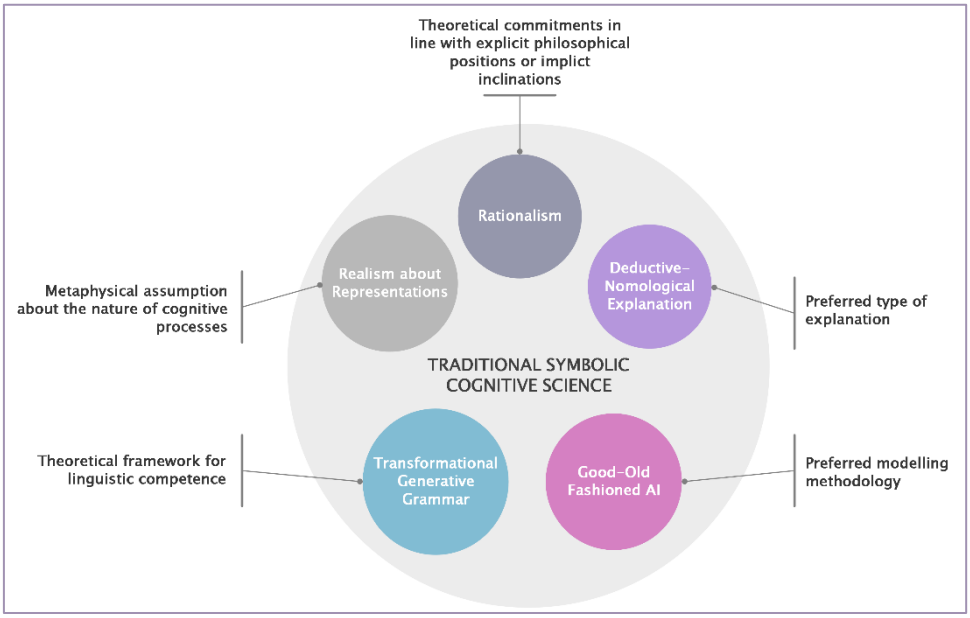


Fig. 1 An overview of traditional symbolic cognitive science

The theoretical framework for understanding the nature of linguistic competence that was incorporated in this strand of cognitive science was transformational-generative grammar (TGG). This framework assumed, along the rationalist line, that our linguistic competence amounts to the innate, domain-specific universal grammar that contains all the rules and principles one needs to master any language, first and foremost their own mother tongue (Chomsky 1957, 1966). TGG understood linguistic competence narrowly as including only syntax, which was taken to be cognitively autonomous, i.e., isolated from other linguistic levels such as semantics, morphology, phonology, or other cognitive processes, such as

sensory-motor processing. Furthermore, most of these influences were quarantined to linguistic behavior or performance as per TGG, thereby creating a gap between idealized competence and idiosyncratic performance. Moreover, TGG singled out essential properties of language, which were then mapped onto thought, thereby making a monolith out of language and thought. The semantic counterpart of TGG, the Language of Thought hypothesis (LOT), stated that both language and thought have semantic content ("aboutness") and are compositional, systematic, and productive, therefore, any research program in cognitive science needed to postulate cognitive architecture that could account for these essential features. This basically means that computational models had to implement manually specified symbolic representations and hardwired discrete rules to be considered faithful simulations of the innate human language faculty that resembles thought in this regard.

Arguably, in the 1980s, connectionist cognitive science, also known as parallel distributed processing (PDP) approach, came to surface and developed as an antipode to traditional symbolic cognitive science almost in all respects. Most importantly, theoretical commitments of connectionist researchers were directly opposed to those of their rivals since they were—and still are—in line with empiricism. The preferred methodology for investigating human cognitive processing is what allowed connectionism to shake things up in cognitive science since artificial neural networks (ANNs) were introduced as a biologically more plausible option than GOFAI in the seminal publication of Rumelhart & McClelland's (1986), AKA "The PDP Bible." However, given that connectionist modelers used non-symbolic vector representations and strived to minimize rules as much as possible, the main line of criticism treated connectionism as either false hypothesis about human cognition or as mere implementation of LOT that has more biological plausibility.

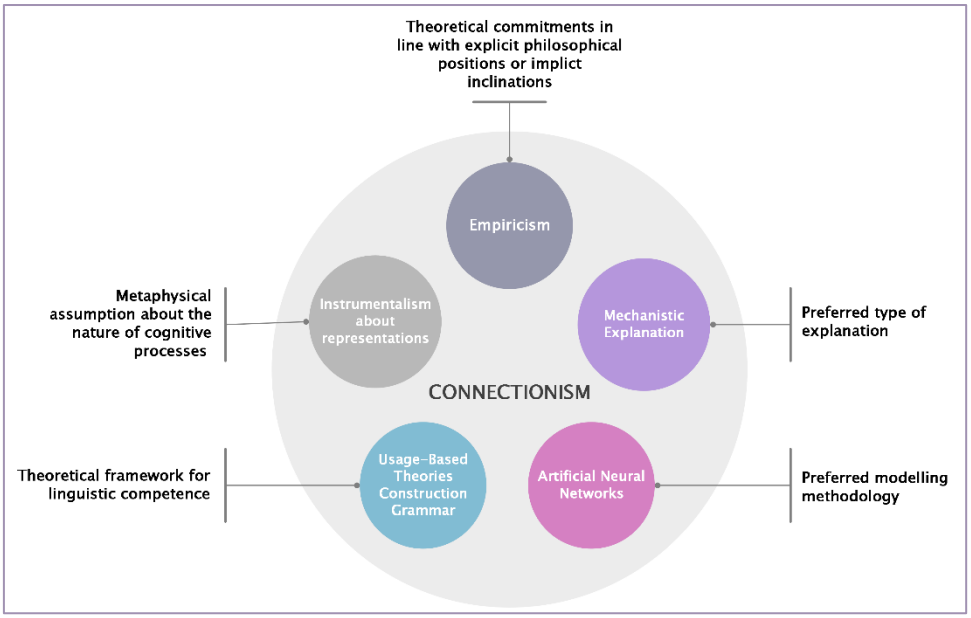


Fig. 2 An overview of connectionism

My intention in the following chapters is to present connectionism as a *unified* research program—described in Fig. 2—that can stand up for itself against traditional

symbolic cognitive science, which has not been done explicitly in the *pro*-connectionist literature so far. Thus, my objective is to defend the banners of empiricism regarding the nature of linguistic competence through connectionist models. I will argue that these models provide us with mechanistic explanations, which makes connectionist contribution to understanding human cognitive processes distinctive. This means that explanatory standards and desiderata inherited from traditional symbolic cognitive science are not adequate for the evaluation of explanatory prospects of connectionism. Furthermore, I will show the compatibility between theoretical frameworks of the Usage-Based Theory (UBT) (Tomasello 2003) and Construction Grammar Approach (CAP) (Langacker 2007), and pattern-based connectionist account of linguistic competence as opposed to rule-based account of TGG. The patterns of linguistic usage are emergent properties of the linguistic performance of connectionist models and cast doubt on the artificially created gap between competence and performance in TGG.

My focus will be on the contemporary connectionist models or post-connectionist models, which differ from classical connectionist models of the 1980s in terms of the number of layers within an ANN, as well as the amount of data they can process, type of algorithm, and architectural features. Post-connectionist models are based on deep learning (DL). DL refers to the algorithm, or learning technique, for deriving an optimal solution to any problem given a sufficiently extensive and relevant dataset (Torfi et al. 2021: 2). In 2016, a DL-based artificial agent, AlphaZero, beat Lee Sedol, the master of Korean Go, which is arguably a more complex game than chess. At that very moment, a bygone syntagma made the press covers—*tabula rasa*. AlphaZero was described as a tabula rasa system that vindicates empiricism by showing that learning from experience and without any innate or manually specified rules results in successful task performance, such as acing the game of Go that requires strategic planning and some sort of creativity (Silver et al. 2017). Thus, we were left with wondering whether DL lives up to the old connectionist ambition of demonstrating that empiricism vs. rationalism debate has an empirically validated winner.

To sum up, I consider proponents of traditional symbolic cognitive science and TGG to be rationalists. On the other hand, I consider proponents of connectionism to be empiricists. Being labeled as a "rationalist" or "empiricist" has to do with the issue of whether linguistic competence is understood as being innate and requiring domain-specific cognitive resources, or as being acquired and requiring domain-general cognitive resources. As I will be discussing in far more detail in Chs. **2** and **3**, these claims are not controversial since leading scientists and philosophers, some of which are enlisted in **Fig. 3**, have explicitly committed to either rationalism or empiricism. The controversy, however, lies in the issue of whether their commitments *really* entail rationalism or empiricism, and what sort of rationalism and empiricism applies to their views. I will argue, by drawing on Cameron Buckner (2018, 2023), that the contemporary strand of the debate has little to do with historical empiricism and rationalism. Rather, the tug of war is about how many priors and inductive biases in DL models can empiricists endorse without dissolving their position into vanilla rationalism. Priors are probabilistic assumptions about the underlying distribution of the data. They represent antecedent knowledge or expectations that are incorporated into the learning algorithm. Inductive biases, on the other hand, are

constraints built into learning algorithms that help guide the learning process by favoring certain hypotheses or solutions over others so that DL models can generalize better or make predictions more efficiently.

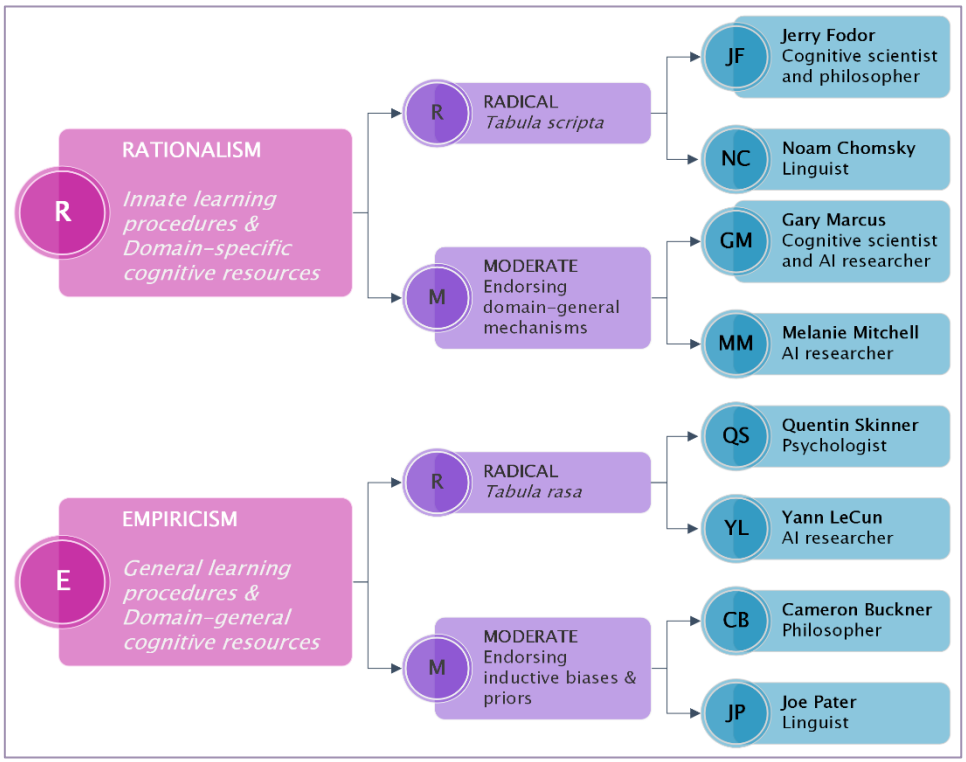


Fig. 3 An overview of philosophers and scientists with respect to their professed position

After the preliminary clarifications regarding terminology, key figures, and the general trajectory of the dissertation, the time has come to set forth the hypotheses that guided the research project I present in the following chapters. The skeleton of the research project was made of one main hypothesis supported by two auxiliary hypotheses and three specific hypotheses. Specific hypotheses are to some extent independent since their role was to establish clear and precise schema of rationalism vs. empiricism debate across centuries and scientific fields, and to investigate to what extent connectionism can be regarded as an autonomous theory about our cognition besides being a valuable tool for modeling cognitive processes. In other words, the issue at hand was to analyze whether we instantiate something akin to connectionist cognitive architecture. Main hypothesis with its auxiliary hypotheses is focused on linguistic competence as the main source of friction between rationalists and empiricists. Defending some of the specific hypotheses (e.g., **I** & **II**) does not have any particular repercussion on the main hypothesis, although some of the specific hypotheses do serve as reinforcement to the plausibility of the main hypothesis (e.g., **III**).

Main hypothesis

If one can show that linguistic competence can be examined, explained and simulated faithfully enough via models of syntactic and semantic processing, which are not based on the application of encoded rules and symbolic representations, but, rather, on DL and huge amount of data, then it is more scientifically fruitful and philosophically convincing to endorse empiricist account of linguistic competence as opposed to rationalist account.[1]

The corollary of the main hypothesis

Rationalists are wrong when assuming that language competence–qua domain-specific faculty–is innate. However, this does not mean that nativism cannot be a viable position when it comes to domain-general mechanisms.

Auxiliary hypothesis A

Contrary to the core assumptions of transformational-generative grammar, syntactic processing is not cognitively isolated from semantic processing. Rather, syntactic and semantic processing are intertwined processes that constrain each other.

Auxiliary hypothesis B

Contrary to the core assumptions of traditional symbolic cognitive science, embodied approaches to cognition should be incorporated in post-connectionist models to account for the dependence of linguistic competence on both body and environment.

Specific hypothesis I

The strong historical influence of rationalist ideas and the Cartesian heritage on the 20th century philosophy of language and theoretical linguistics can be detected. This influence stretched to cognitive science thanks to Noam Chomsky and is reflected in the ontological assumption that there is correspondence between language and thought regarding the allegedly essential properties such as systematicity and productivity.

Specific hypothesis II

Labeling a philosopher or a scientist as a rationalist or empiricist in the 20th and 21st century has a different connotation than it had in the history of philosophy because it is dependent on the additional theoretical commitments that linguists, cognitive scientists, and AI researchers implicitly or explicitly assume.

Specific hypothesis III

Connectionism is autonomous qua theory of human cognition and approach to modelling human cognitive processes – contra criticism stemming from the traditional symbolic cognitive science – because it provides us with theoretical and computational means to decouple language from thought, thereby opening the possibility of the simultaneous existence of the systematic language and non-systematic thought.

---

[1] Here it is important to note that a conjunction constitutes the antecedent. With this conjunction, the antecedent either "collapses" or "survives", because, if it had been formulated in a less committing way, then the hypothesis would have amounted to a mere instrumentalist position. The way I formulate the consequent allows me to avoid the rationalist attack that would revolve around the allegedly firm scientific and philosophical basis of innateness. As per my hypothesis, such a move would be a typical example of logical fallacy of negating antecedent and consequent.

Each of the hypotheses will be tackled within the four chapters of this dissertation, most often more than one hypothesis per chapter. The first two chapters are significantly shorter than the third and serve as a praeludium for introducing state-of-the-art DL models and LLMs that represent the axis of the new account of the linguistic competence I will be defending in the fourth chapter. Thus, these chapters should be read as a broader historical or theoretical context for understanding intricacies related to the competing views of linguistic competence. The consequences of **Specific hypotheses II** and **III** will be made clear in Ch. **3**, thereby rounding up the argumentative line stretching from Chs. **1** & **2**, where its development, along with **Specific hypothesis I**, got off the ground. The crux of the dissertation is Ch. **4**, in which I strive to analyze and defend the **Main hypothesis**, corollary, and **Auxiliary hypotheses A** & **B.** Hence, the condensed philosophical and scientific jargon, tedious technicalities, and elaborate arguments reside in Chs. **3** & **4** since the aim is to offer a novel account of linguistic competence inspired by LLMs and entrenched in the new empiricist dogma as advocated by Buckner (2023), that can be considered as the fortress from which empiricist currently hold rationalists at gunpoint. In **Conclusion**, I sketch the philosophical implications stemming from the novel account of linguistic competence, especially for subfields such as (analytic) philosophy of language and mind, whose foundations rest on the endorsed view of the nature of our language capacities. In the roots of rationalist philosophy of language and mind, as well as theoretical linguistics in transformational-generative tradition, is the conviction that language makes us unique, i.e., sets us apart from the rest of the mammals, and ensures our privileged status despite the proliferation of artificial agents, such as chatbots, that mimic our behavior.

The relevance of this dissertation is best seen in the light of the recent breakthrough in conversational Artificial Intelligence (AI). In the past couple of months, digital media, social media, and traditional broadcasting media are all buzzing with the word "ChatGPT." The word designates a state-of-the-art chatbot (OpenAI 2022). Conversational AI has been around for fifty years—ever since ELIZA, the first AI therapist implemented into a GOFAI model, asked a human being "Is something bothering you?" (Weizenbaum 1966). However, this time, the fuss was different. Unlike ELIZA, which had encoded script titled DOCTOR to follow, ChatGPT was trained on 570 GB of textual corpora to learn how to interact with us in more than 95 world languages. Most importantly, ChatGPT is based on DL and thus considered as an implemented large language model (LLM) GPT-3, which stands for generative pre-trained transformer (Brown et al. 2020).[2]

A heated debate has ensued in which academics of all kinds discuss whether ChatGPT understands the meaning of sentences it produces (Shanahan 2023, Durt, Froese & Fuchs 2023). Three years ago, when I started developing the structure and argumentation of the dissertation, the success of DL was evident in domains such as computer vision, but for natural language processing (NLP) it still seemed far-fetched. This was not surprising

[2] Meanwhile, on March 13th, 2023, GPT-4 was released, with a reported 100 trillion learned parameters, which is circa 571 times as many as for GPT-3. The amount of training data is still unknown. The difference between GPT-3 and GPT-4 is in the type of data: whereas the former is trained only on textual data, the latter is trained on images as well. In other words, GPT-4 is a clear example of a multimodal LLM.

at all, given that DL originates from connectionist models which became prominent in the 1980s and were great for simulating lower cognitive processes such as perception, but notoriously bad for simulating higher cognitive process such as language. It seemed that only a skosh bit of papers expressed optimism regarding prospects of DL when it comes to NLP. But then, as of November 2022, everyone is worrying to what extent a DL-based conversational AI is sentient, ready to take jobs from professional translators, writers, and editors, or jeopardizing the entire system of grading students' essays.

My dissertation examines whether DL models, including LLMs, are informative regarding the nature of linguistic competence and to what extent they vindicate empiricism about linguistic competence. Thus, witnessing the genesis and development of one of such models, such as, GPT-3 within ChatGPT, which exhibits remarkable and anomalous linguistic behavior at the same time, could not be a better testbed for philosophical points and arguments expressed here. What a time to be alive, right? Obviously, the dissertation rests on endorsing many mutually compatible but anti-mainstream positions, and for this reason, some of the points and arguments may be too bold (after all, they don't say "young and bold" for nothing), or *prima facie* doomed for an attentive armchair analytic philosopher, zealous nativist cognitive scientist, or linguist. To them I dedicate the immortal words of one of the greatest rationalists, Jerry Fodor, beneath the title of this Ch.

1. LINGUISTIC COMPETENCE: A VERY SHORT HISTORY

> *Now the Lord God had formed out of the ground all the wild animals and all the birds in the sky. He brought them to the man to see what he would name them; and whatever the man called each living creature, that was its name. So, the man gave names to all the livestock, the birds in the sky and all the wild animals.*
>
> —*Genesis* 2:18-20

1.1. The Pre-Cartesian Era: The Overlooked Roots

Most histories of linguistics begin with the 19th century, often ignoring its philosophical legacy. On the other hand, most intellectual histories of virtually anything in philosophy begin with antiquity. Either you have a penchant for Plato or Aristotle. Either you are looking up to the sky in search for the perfect and ever-lasting Ideas or you are inspecting the earth in search of the fellow featherless bipeds. Plato was looking for the reality in which Ideas manifested to the mind's eye, Aristotle was more interested in the instantiated Ideas, i.e., mind-independent materialized forms. The first inaugurated the search for the underlying meaning, the second introduced logic as the instrument for reasoning and classifying everything that was deemed to exist. It would not be entirely wrong to call them proto-rationalist and proto-empiricist. My brief history of linguistic competence will start even earlier and then fast forward to much later. Given that I am not a historian of ideas, nor is this dissertation a piece in the history of philosophy, I am not following a linear timeline or linear development of ideas by listing philosophers one by one. Rather, my aim is to trace the origins of the divided views of linguistic competence specifically and means to understand human cognitive processes generally. The divided views revolve around the question whether cognitive processes, such as language, are innate or acquired. By far the most influential view of innate linguistic competence was put forward by Noam Chomsky, who dedicated the whole book to pinpointing the historical roots of his TGG. My brief historical overview should be read as an addendum and comment on Chomsky's *Cartesian Linguistics* (1966) since I will be discovering quite thought-provoking gaps in his historical overview *and digging* deep on the other side of the trench to build defenses for empiricism, according to which experience is more important for understanding linguistic competence than alleged innateness.

Thus, my preoccupation will be to address **Specific hypothesis I** which states that one can detect rationalist influence on the 20th century philosophy and linguistics, as well as on cognitive science, which is grounded in the ontological assumption that language and thought are isomorphic with respect to their essential properties. This is why my historical overview will be more focused on shedding light on the theoretical commitments of the adversarial side—know thy enemy, as Sun Tzu would say. Along the way, I will also introduce key figures of empiricism. In both cases, I will cherry-pick the least controversial authors with respect to their commitments—they are quite explicit in their argumentation.

I will also touch upon **Specific hypothesis II** only to continue its defense in the next Ch. Recall, this hypothesis stresses the change in labeling: being either rationalist or empiricist in the 20th century is not quite similar to adhering to such positions in the 21st century.

Our story begins in the Garden of Eden. If anything, the Christian and Hebrew Bible (the New and Old Testament along with *midrashim,* i.e., Rabbinic interpretations) is a cornerstone for the Western civilization, and in it, the origins of linguistic competence, as well as the origins of humankind, are subsumed under the aptly named part *Genesis*. Long story short, God creates everything, then proceeds to creating Adam (and Eve), and bestows upon Adam the capability to name everything that was created, most notably animals. This sacred language, that both God and Adam spoke in the Garden of Eden, was later dubbed the Adamic language (Eco 1995: 7).[3] From there, everything went wrong. Adam (and Eve) were expelled from the Garden of Eden due to eating forbidden fruit and we ended up with *confusio linguarum*, i.e., a bunch of imperfect languages. Luckily for us, the God kindly reminded Moses of the following when Moses hesitated to ask Pharaoh to let his people go: "Who has made man's mouth? Who makes him mute, or deaf, or seeing, or blind? Is it not I, the Lord?" (*Exodus* 4:11). In other words, it is God who made us capable to speak, hear, and see, and this is what makes us unique, despite the crumbling of the Adamic language. This also allows us to dedicate our capabilities to the quest of uncovering and recovering the Adamic language.

Fast forward to the Middle Ages, specifically the 14th century. William of Ockham, a Franciscan friar and philosopher from British Isles, well-versed in Aristotle's logic, put forward the idea of mental language (*oratio mentalis*) to which Fodor's LOT and Chomsky's TGG bear a striking resemblance although neither one of them was acquainted with Ockham's work (Normore 2009: 293). Some philosophers (e.g., Geach 1957) were inclined to think that Ockham's mental language was conceived as Latin, i.e., what we say on the inside is what would be otherwise said aloud in Latin; while others (e.g., Trentman 1970, Nuchelmans 1992) have seen in it a perfect Adamic language in which there would be no need for synonymy or equivocation.[4] What needs to be cleared at once is the very notion of mental language. For Ockham, mental language is prior to our mundane mode of communication relying on conventional signs and it is shared among all rational beings since it relies on natural signs (Panaccio 1999: 53). Being a nominalist and a fervent supporter of the principle, later called Ockham's Razor, he believed that there are only singular entities in the world. Thus, mental language, in virtue of providing the grounds for conventional, external language, links concepts to singular entities.

The process of linking proceeds via signification, a primitive term of Ockham's semantics (Normore 1990: 54). This signification is a conventional one, but there is also

[3] This language may or may not be Hebrew—Dante Alighieri believed to be so (Latin being only an artificial homologue), and in the Middle Ages there were many apocryphal stories about children who would automatically speak Hebrew despite not being exposed to any sort of speech or linguistic stimuli (Eco 1995: 33-35). Leibniz, nonetheless, believed it was German rather than Hebrew (Aarslef 1982a: 46).

[4] See Spade (1980) for casting doubt on Trentman's (1970) arguments that mental language allows for no synonyms or equivocations. Spade finds the inconsistencies in Ockham and maintains that he did not fully develop mental grammar (i.e., syntactic aspect) because he focused too narrowly on truth conditions.

natural signification, which specifies the origin of mental terms *per se*.[5] Unlike contemporary nativists, such as Fodor and Chomsky, Ockham did not adhere to the innate vs. acquired dichotomy. Natural signification is causal: encountered objects "produce" mental terms as effects in a soul, and in this sense mental terms are acquired even though they are not learned (Normore 1990: 56). Taking signification as semantic fuel, the suppositions, or propositions, start to form, and each determines the domain of objects of reference to which terms will be applied (Panaccio 1999: 59). Propositions are functions of terms that constitute them, much like in Fodor's LOT molecular representations are formed out of atomic ones via discrete rules.

The mental language encodes semantics for conventional languages almost like LOT encodes semantics for natural language: the "deep structures" of mental language offer us the means for expressing truth conditions that underly diversity and plurality present in the "superficial structures" of conventional languages (Nuchlemans 1992: 50). The famous chasm that will divide analytic philosophers of language—the difference between extension (i.e., a term that designates an object), and intension (i.e., meaning or sense of the term)—has also roots in Ockham, as Nuchelmans (1992) rightly remarks. The s*alva veritate* substitution is a useful tool for distinguishing the intensional from extensional contexts: in the former context, co-referring terms are not interchangeable, but in the latter they most certainly are. Thus, some terms of mental language cannot be substituted *salva veritate* with conventional terms since mental terms have a single meaning, while conventional terms can have different roles within sentences of natural language, i.e., functional roles which make a noun out of a single term in some contexts and adjective in other. This is why conventional terms are always subordinate to mental terms: mental terms are never equivocal whereas conventional terms are. On the other hand, *salva veritate* substitution is feasible if and only if the two conventional terms correspond to the same mental term because then they can be said to have the same signification given that each mental term has exactly one signification. As Normore (1990: 55) convincingly argues, if one changes "mental term" with "sense" or "*Sinn*" and "signification" with "reference" or "*Bedeutung*", one gets a proto-Fregean theory of reference four centuries earlier (a teaser for Sect. **1.3.**), albeit in the Aristotelian logical framework rather than symbolic.

In history, however, silence and absence are sometimes more interesting than voice and evidence: the mental language vanished from philosophical argot and sources by the end of the 14th century and did not re-emerge until the 20th century and Fodor's LOT. Fast forward to the high Renaissance, specifically the 16th century. Renaissance scholars generally harnessed a certain despise towards logic considering *usus loquendi*, viz., customary speech, as more valuable since it sheds light on philosophical and scientific inquiry while logic obscures it through empty technicalities (Losonsky 2006a: 183).[6] In line with the trends of his epoch, Francisco Sánchez de las Brozas, also called Sanctius, a

[5] Furthermore, signification can be primary (both *sensu stricto* and *sensu lato*) and secondary, or connotation to enhance the ontological economy in order to avoid introducing the synonymy. For details see Panaccio 2006: 56-58.

[6] Of course, this is only one of the possible reasons for the disappearance of mental language à la Ockham, for an array of other reasons *cf*. Normore 2009.

professor at the University of Salamanca, wrote a manual on Latin grammar titled *Minerva seu de Causis Linguæ Latinæ (*better known as *Minerva*) in 1587 (Seuren 1998: 42). The manual was written in the Renaissance spirit of preferring natural language over Aristotelian syllogistic, hence Sanctius put an emphasis on syntax, as opposed to semantic concerns of medieval logic. In Pieter Seuren's words, "we thus have here a precursor of transformational grammar (...), though in less modern terminology."

Sanctius acknowledges the datum of the Aristotelian logic—that influenced medieval logicians like Ockham—that there must be a correspondence between logical categories and structures of thought as well as between structures of thought and the world (Seuren 1998: 45). Both kinds of correspondence hinge on language since we express thoughts through language. However, Sanctius realized, again much like Ockham, that *confusio linguarum* and semantic anomalies, such as equivocacy, indicate that sentences in our everyday language, i.e., their *surface structure*, cannot yield the two kinds of correspondence. Something more perfect is needed, an abstract level on which sentences and thoughts map one to one, i.e., one must find the *deep structure*. The grammar of any language provides the rules for transforming surface structure into deep structure. At this point, Sanctius was already original enough, but he went even further in analysis: specific grammars combine into universal grammar at even more abstract level because languages are in principle translatable to each other, and virtually every group of people speaks some language, otherwise they could not communicate with each other. This universal grammar perfectly fits the structure of thought and reflects the features of original language that God bestowed on Adam because this would be the ultimate mirror of our rational capacity that sets us apart from animals (Seuren 1998: 65-66). To sum up, every human being is endowed with universal grammar as mark of God and in virtue of being rational. Sanctius was not explicit about the innateness of such a syntactic device, but rather outsourced it to God. Thus, excluding the religious narrative accompanying it, the universal grammar of Sanctius bears a remarkable resemblance to Chomsky's, which will be presented in Sect. **2.1.**

The *Minerva* had a peculiar destiny after the death of its author. After being ignored for almost a century, it was rediscovered by one of the key figures of Port Royal monastic intellectual milieu, Claude Lancelot, who integrated the ideas of Sanctius into Port Royal grammar to the extent that *Minerva* ceased to exist without Port Royal flavor (Aarslef 1982b: 104, Seuren 1998: 47). This was chosen as the starting point of Chomsky's *Cartesian Linguistics* rather than any medieval or renaissance intellectual authority. However, my aim for this Sect. was to show that both Ockham and Sanctius can be seen as precursors of the main research questions that I will be tackling within this dissertation: *Is the thought structured like language, or it may diverge from linguistic structure altogether*? *Is it necessary for the cognitive architecture that underlies thoughts to incorporate properties corresponding to the properties of natural language?* To put it differently, do the nature of thought and the nature of linguistic competence coincide? Are thinking and language processing coextensive? This research question is also at the core of TGG and LOT, each being preoccupied with the domain of linguistic competence that seemed prone to more exact treatment—TGG with what I will baptize syntactic competence and LOT with semantic competence (Sect. **4.2.** & **4.3.**). This aligns well with Ockham's focus on semantics and Sanctius' focus on syntax.

1.2. Early Modern Rationalism and Empiricism: A Gap between Reason and Experience, Continent, and the Island

Philosophers like polarities and grouping into camps, which often makes their discussions similar to pep rallies. One of the perennial divisions that nowadays has progressed well beyond disciplinary corners of philosophy is the division into rationalist and empiricist coterie, whose origins date back to the 17th century. In this brief historical overview, I will touch upon Descartes, his Port Royal successors, and Leibniz, i.e., the central figures of rationalism writing about linguistic competence, and Locke, the central figure of empiricism who dedicated a book of his *Essay concerning Human Understanding* (1690) to natural language thereby marking the beginning of the philosophy of language as we know it (Losonsky 2006b). A more extensive treatment of this fecund period of the history of philosophy certainly merits a dissertation of its own.

Descartes, (in)famous for his dualism, had little to say about natural language, since the linguistic capacity was, for him, subordinated to thought, which is the essential attribute of *res cogitans*, viz., thinking thing. One of the particularly relevant (and long) passages is the following:

> "For it is a very remarkable thing that there are no men, not even the insane, so dull and stupid that they cannot put words together in a manner to convey their thoughts. On the contrary, there is no other animal, however perfect and fortunately situated it may be, that can do the same. And this is not because they lack the organs, for we see that magpies and parrots can pronounce words as well as we can, and nevertheless cannot speak as we do, that is, in showing that they think what they are saying. On the other hand, even those men born deaf and dumb, lacking the organs which others make use of in speaking, and at least as badly off as the animals in this respect, usually invent for themselves some signs by which they make themselves understood. And this proves not merely animals have less reason than men but that they have none at all, for we see that very little is needed to talk" (CMT 1 140).

This passage picks out all crucial ideas of rationalism when it comes to language: it is a capacity unique to humans and reserved for humans due to the essential property of creativity, thereby making them special in the natural order of things. This is what appealed to Chomsky (1966): creative usage of language that defies practical purposes points to productivity as essential property of both language and thought. Descartes' dualism also hints at the difference between surface and deep structure: vocalization of particular languages has to do with *res extensa* while the true nature of any language has to do with *res cogitans*. Moreover, the Cartesian linguistic capacity is intrinsically linked to reason: mind without language would, in fact, inhibit reason (Losonsky 2007: 185). And the reason is full of innate ideas, for instance, those of God (CSM 2 35) and mathematical principles (CSM 2 262). However, not only the existence of God and eternal truths is innate, but elsewhere Descartes endorsed universal innateness according to which even our sensory ideas are innate.[7] Descartes thus introduced nativism as the rationalist building block for

[7] See Gorham (2002) for an extensive treatment of the causal and non-causal interpretation of the universal innateness thesis in Descartes. I remain neutral towards either interpretation since it is only relevant whether language is innate for the purpose of this Sect.

the epistemology, although ultimately the innate faculty of reason was made possible by a benevolent God "concerned to give us a head start in our attempts to negotiate the sublunary wilderness" (Cowie 1999: 9).[8]

Inclined to some of the tenets of Cartesianism, a group of Jansenist intellectuals at Port Royal, most notably Claude Lancelot who got hold of *Minerva*, Antoine Arnauld, and Pierre Nicole published *Grammaire générale et raisonnée contenant les fondemens de l'art de parler, expliqués d'une manière claire et naturelle* (or simply Port Royal *Grammaire*) in 1660 and *La logique, ou l'art de penser* (or simply Port Royal *Logique*) in 1662. The books are intertwined given that their hypothesis was that logical operations of mind give rise to grammatical features of different languages (Seuren 1998: 47). However, given that there is only one true logic, namely Aristotelian terminist logic embodied in syllogistic, all languages must be somehow related, for instance via deep structure that is represented through one-to-one mappings between constituent structures and thoughts. As Lancelot & Arnauld say in Port Royal *Grammaire*, to grasp the universal language comprised of deep structures "it would be enough to examine thoughts in themselves, unclothed in words, or other signs" (cit. in Losonsky 2006a: 186). Obviously, Lancelot & Arnauld proceed with the Cartesian subordination of language to thought and distinguish deep from surface structure in the sense that language for them has universal syntactic skeleton along with the spoken and written letters which are not necessary for understanding the foundations of grammar but merely serve for communication. Without assuming that there is an abstract, deep linguistic *substratum*, we would be left with cultural relativism and common usage which is unstable and uncertain. And we cannot be uncertain about something that was a gift from God, can we? Thus, Chomsky encircles his 17th century rationalist and nativist pedigree with Port Royalists.

Nonetheless, some people were not ready to acknowledge this pedigree. I have already shown the significant gaps in the Cartesian story about the rationalist origins of linguistic competence, namely the role of both Ockham and Sanctius. Additionally, the search for universal grammar did not encompass only syntax but also semantics. Furthermore, Port Royalists did not really see eye to eye with Descartes regarding his thesis of universal innateness (Aarslef 1982: 104), while Chomsky seems to merely lump together everything. As Hans Aarslef puts it: "Professor Chomsky has significantly set back the history of linguistics. Unless we reject his account, we will for a long while have no genuine history, but only a succession of enthusiastic variations on false themes" (1982b: 116-117).

[8] Fiona Cowie (1999) adequately distinguishes epistemological from psychological questions to which rationalism/nativism should provide an answer and shows that 17th century rationalism when dealing with the epistemological question relies on God for justifying knowledge and, specifically, a priori beliefs. However, when the problem at hand is the very origin of knowledge and such beliefs, rationalism dissolves into nativism as a main strategy for tackling the psychological question. In this Ch., and in the dissertation generally, I am obviously focused on the psychological question rather than epistemological, and even more narrowly, on the origin of linguistic competence. However, Cowie (1999) is a *locus classicus* for an extensive take on both questions from the historical *and* contemporary perspective, as well as for a wholesome criticism of rationalism/nativism in this regard. In the 21st century a book of similar importance and scope is Clark & Lappin (2011) who disagree with Cowie's approach and insist that the boundaries between rationalism and empiricism should be draw according to their allegiance to metaphysics (rationalists) or experimental science (empiricists).

And, alas, the messiness of the contemporary rationalist vs. empiricist debate across cognitive science, linguistics, and AI research corroborates Aarslef's grim prediction to a great extent (teaser for Sect. **3.2.**). According to Aarslef (1982b: 106), to be a universal grammarian, it suffices to be rationalist, and I would add that it suffices to be both rationalist and oriented toward language of logic rather than natural language. The logic will change in time, from Aristotelian syllogistic system to the first and second-order logic of Frege and Russell, but the obsession to mold natural language and thoughts into it will not. In the past, universal grammar was seen as glimpse of the Adamic language, to which logic can bring us closer, whereas in the present, universal grammar is a supposedly innate device to whose modelling logic can bring us closer.

Leibniz was one of the fervent supporters of the idea that logic is a perfect tool for unraveling the true nature of relationship between thought and language and, thus, a source of inspiration for the logicians and philosophers of language in the 20th century. Leibniz was also interested in the Adamic language and immersed himself in the studies of etymology, most notably etymology of German, French, and Slavic languages. The diachronic perspective allowed Leibniz to claim that there is a natural order in words' origins as though they are all converging to a common point. This common point must be a single language like the Adamic language, which keeps the concealed truth and wisdom since it was common to humans and angels alike (*New Ess*. III.ii.1, *cf*. Aarslef 1982a: 59).

The only thing that could be on a par with the Adamic language is a formal language decluttered from ambiguities and other semantic anomalies since both can share structure and preserve truths much like decimal and binary system preserve truth about natural numbers (Losonsky 2006a: 192). For this reason, Leibniz devised *Characteristica universalis*, a precursor to the notation that Frege will introduce for modern symbolic logic in the 20th century. Thus, by iteratively performing substitution *salva veritate*, one could, in Leibniz's view, transform sentences of any natural language to formal language of *Characteristica universalis*, which, in turn, reveal the prefect logical form of such sentences. This aligns with the points of Sanctius and Port Royal group, as well as their emphasis on universal grammar, i.e., it fits the recurring rationalist template. More importantly, however, Leibniz *qua* rationalist disagreed with Locke on virtually every aspect of their respective treatment of language and linguistic capabilities except for the starting point that language is indeed a suitable instrument for looking inside the human mind. Leibniz saw in it the inner deductive structure of the mind that reflects the grammar of natural language synchronically and the grammar of Adamic language diachronically, while Locke saw psychological cues about the natural language and contents of thoughts prompted by sensory stimuli. And thus, the seed of discord was planted.

Locke, the father of British empiricism, linked semantics to epistemology, which was unprecedented in the history of philosophy before him (see Losonsky 2006b). The degree of human knowledge, for him, hinges on the manner of expressing it, and if one strives to chart the origins of human knowledge, she must begin with language (of course, this being after the primary sensory input). As Locke put it, humans, in fact, "in their Thinking and Reasoning with themselves, make use of Words instead of Ideas" (*Essay*, IV.v.4). This is because the words "in their primary or immediate signification, stand for nothing, but the

ideas in the mind of him that uses them" (*Essay*, III.ii.2). Panaccio (2003) uses this sentence to show a straight line between Ockham's nominalism and Locke's view on words. Hannah Dawson claims (2007: 187-188) that in this seemingly clear and simple formulation the roots of Locke's rebellion against the tradition are to be found: subjective, sensible words of the individual *speaker* come to the forefront rather than mental propositions. The concerns pertaining to the meaning of words are no longer generic, nor is the solution to such concerns generic and abstract: the concrete individual utters the words of an imperfect natural language ("vulgar speech") *qua* contingent phenomenon.

Interestingly, however, Locke did not intend to dedicate a whole book to the issue of language but given that his *Essay* was written over the course of twenty years, it took him time and effort to supply it with a treatment of any novel idea that appeared during the period of writing. Thus, Aarslef (1982a: 45) draws on the published lists of Locke's travel literature and books sent home from France during his travels to claim that Locke, in fact, read Port Royalists and found them so compelling that he devoted a whole book to discarding the rationalist account of language. Unlike Port Royalists and Descartes, he was not swept away with either logic or mathematics, nor searching for the perfect non-contingent Adamic language. Locke denied the authority of Adam based on his first baptismal act by reckoning that "[t]he same liberty also, that Adam had of affixing any new Name to any Idea; the same one has any one still (especially the beginners of Languages...) but only with this difference, that in Places, where Men in Society have already established a Language among them, the signification of Words are very warily and sparingly to be alter'd..." (see *Essay* III.vi.43-51).

Locke does not sugarcoat the fact that language is full of semantic anomalies, ambiguities, and that the ephemeral and conventional nature of language makes it inherently unstable—after all, every speaker has something to say based on the ideas she previously entertained. The core issue here is that the speaker, according to Locke (*Essay* II.xii.1), entertains complex ideas rather than simple ones since the former are actively formed, while the latter are passively received through senses. And while all human knowledge begins with senses as *tabula rasa* becomes marked with some chalk, complex ideas such as those of number, duration, causation, or space must be combined by psychological capacities out of simple ideas and this is chiefly a rational process. The process of combining simple ideas into complex ones fluctuates with respect to intra and interindividual differences, which Dawson (2007) dubs intra and interpersonal semantic multiplication. Semantic multiplication is mediated through categorizing, i.e., the formation of more and more abstract labels for perceived objects that figure in semantic content in virtue of being ideas. Abstract label does not, however, imply the existence of an abstract idea. At this point, Locke invoked the example with triangle which marked the era of British empiricism and will even figure prominently in contemporary vindication of empiricism within connectionist cognitive science (Buckner 2018). Here is the passage:

"For when we nicely reflect upon them, we shall find, that general Ideas are Fictions and Contrivances of the Mind (...) For example, Does it not require some pains and skill to form the general Idea of a Triangle, (which is yet none of the most abstract, comprehensive, and difficult,) for it must be neither Oblique, nor Rectangle, neither Equilateral...; but all and none of these at once. In

effect, it is something imperfect, that cannot exist; an Idea wherein some parts of several different and inconsistent Ideas are put together" (*Essay* IV.vii.9).

In other words, only a few exemplars can be subsumed under a general name or abstract label, and virtually every single speaker can treat the meaning of such a label differently, which, in turn, has practical consequences. This is what impedes our morality given that people disagree about what is morally right or wrong unless one calls upon God as the ultimate arbiter. Alas, this usually just muddies the waters since having justified beliefs is not the same as putting faith into God, which essentially blunts our critical and moral capacities because we start acting like children who were merely instructed to behave properly (Woolhouse 1988: 95). In sum, with nothing to anchor semantic multiplication since the innateness and abstract ideas are out of question, we are left with the thoroughly arbitrary state of affairs—as Dawson nicely puts it, "Locke's thoroughgoing rejection of innate ideas makes our minds the subjects of whatever chance puts in our path" (2007: 231).

Moreover, the issue of the viability of abstract ideas proved quite challenging for British empiricism in years to come. Hume offered a more nuanced distinction between contents of our minds by introducing impressions, i.e., first appearances in the soul, and ideas *à la* Locke, i.e., faint images of impressions in the soul, as well as associations through which ideas combines and impressions mix (T 1.1.1.1/1). Ideas are copies of impressions and represent them faithfully enough since they represent literal objects that caused impressions (T 1.1.1.7/4). It follows that, should the abstract ideas exist, they could not be traced back to impressions. Whereas Locke thought that we merely subsume exemplar under an abstract label, Hume argued that all we have at the beginning are impressions, thus, we are, in fact, always imagining a concrete exemplar rather than abstract idea. However, the manner in which we select exemplars is puzzling:

"...the seemingly 'picked out' ones—the very ideas […] are thus collected by a kind of magical faculty in the soul. This faculty is always most perfect in the greatest geniuses…but it can't be explained by the utmost efforts of human understanding" (EHU I.7).

Anyhow, let us go back to the Continent and away from Albion. Leibniz, having read the *Essay,* tried to establish correspondence with Locke, but these were futile efforts. Locke's silence resulted in Leibniz's *Les Nouveaux Essais sur l'entendment humain*, in which two characters, namely Philaleth ("truth lover" or "truth enthusiast") and Theophile ("God lover" or "God enthusiast"), debate—as Leibniz expected to debate with Locke in person—about the nature of knowledge. Leibniz started with claiming that ideas *à la* Locke are essentially actual or occurrent thoughts and they indeed capture contingency of human psychological processes, but the mind's underlying logical structure shows the very possibility of entertaining *any* thoughts through dispositions (*New Ess*. III.v.3). Hence, certain classes of ideas and principles, most notably necessary truths that are derivable from dispositions, are innate even though experiential input needs to trigger the mind for the acquisition of many others. Our role is not to roam over the earth by chance and wallow in *confusio linguarum*, despite "[t]he fact is that our needs have forced us to leave the natural

order of ideas (...)" (*New Ess*. III.i.5).[9] Rather, we should discover the pre-established harmony enveloping humans and nature alike: God endowed us with innate principles that correspond to principles governing nature, thereby creating a perfect mechanism not a grab-bag of contingent features. Our rational soul resonates with the rest of the God's creation. After all, there must be a sufficient reason for everything.

Finally, let me state what empiricism and rationalism imply *qua* historical positions involved in *a priori* debate about the origins of knowledge (*cf*. Woolhouse 1988: 2):

> (**E**$_H$) The contents of human minds, i.e., knowledge, and capacities constituting the human minds are grounded in sensory experience and to the great extent acquired rather than innate. Experience is also a touchstone of meaning, truth, and/or any abstract notion whatsoever.
>
> (**R**$_H$) The contents of human minds, i.e., knowledge, and capacities constituting the human minds are shaped by our innate rationality with which God endowed us and thus made us unique in the nature. This is also a touchstone of meaning, truth, and/or any abstract notion whatsoever.

Thus, while the latter camp seeks to establish a rational deductive system of the world and relies on logic as being the tool and ideal to which natural language should aspire, the former camp relies on common usage, common sense, common knowledge and observation. As Plato and Aristotle on *maestro* Raphael's famous painting, the proponent of **R**$_H$ points up to the heavens above and the proponent of **E**$_H$ humbly looks to the soil and immediate surrounding: while one pursues the lost innocence and perfection of the Adamic language, the other drowns in the *confusio linguarum*, trying to make some sense of it. And, as the story continues, the painted Plato and Aristotle just obtain more shadows behind them, as I will be revealing below in the case of analytic philosophy of language.

1.3. Rationalism and Empiricism of the 20th century: Philosophy of Language on the Battlefront

So far, we have seen the unleashed rationalist beast, fed by the Aristotelian syllogistic and empiricist knight waving a sword of natural language grammar. Curiously, however, Aristotelian syllogistic was also embedded in natural language—only Leibniz realized, during his work on *Characteristica universalis*, that conjunction and disjunction of terms in the Aristotelian logic could be somehow represented by arithmetic operations of addition and multiplication (Kneale & Kneale 1962: 404). He lacked formal and notational means to express this relation that will be cornerstone of modern symbolic or mathematical logic. Fast forward to the 19th century and the very beginnings of the new logic. In 1847, George

[9] According to Aarslef (1982a: 69), Locke's *Essay* incited more than philosophical worries in Leibniz: he regarded *Essay* as the proof of alarming impiety that could easily slip into materialism, i.e., denial of both immaterial soul and Christian conception of the life after death. Hence, this probably gave Leibniz the impetus for his *Nouveaux Essais* conceived as pious philosophy on the Continent as opposed to the profanity of decadent Isles.

Boole achieved what Leibniz strived for and published his method in the *Mathematical Analysis of Logic,* the method being the application of algebraic formulae for expressing Aristotelian syllogistic (Kneale & Kneale 1962: 407-412). Thus, "All men are mortal" became $xy=x$, where multiplication of x with y amounts to conjunction of sets or classes. Additionally, the propositions are assigned with truth values, which are represented with either 1 ("true") or 0 ("false"), and Boole goes on to show how anything that can be represented in the algebra of 1 and 0 can be restored within the algebra of classes (Kneale & Kneale 1962: 415). Subsequent development of Boole's new logical method included polyadic predication as well as universal and existential quantification over predicates. Unlike Aristotelian terminist logic, which presupposed that terms have the most important role since they make up syllogisms, the new logic was oriented toward larger units— classes, propositions, quantified expressions.

The time was ripe for Gottlob Frege to enter the stage. He made two long-due improvements. First, he organized the contributions of his predecessors by creating a unified notation through which the structure of logic would be clarified. The formal rigor, precision, and the elaborate deductive system were needed to show that all branches of mathematics can be reduced to arithmetic, which includes only concepts from logic. Thus, his second improvement amounted to finding means to reduce the very process of deduction to a few rules "so that there may be no danger of our unconsciously smuggling in what we ought to prove" (Kneale & Kneale 1962: 436). My focus here will not be Frege's revolutionary approach to mathematical logic and philosophy of mathematics, but rather his philosophy of language. In his *Begriffsschrift* (1879/1972), the primary goal was to set logic free from its link to natural language that was forged in the Aristotelian logic: the relation of his newly envisaged notation, or concept-script, to natural language is akin to the relation between microscope and the eye. In other words, Frege wanted to idealize away from all peculiarities, anomalies, and ambiguities of natural languages, i.e., anything that would hinder one-to-one correspondence between sentences in formal language of propositional and predicate calculus and truth-conditions. Harris (2017) dubs this truth-conditional idealization and argues that both the groundwork for the truth-conditional semantics and the tendency to put an emphasis on declarative sentences can be traced back to Frege. This was the tradition that mostly ignored the richness of natural language, including interrogative and exclamative sentences, not to mention sociolects and idiolects.

Frege's philosophy of language does not hinge on the context sensitivity of natural language or speakers' communicative practice, but is aligned with his broader mathematical goal, and requires, therefore, the same standard of formal rigor and precision. Thus, no longer the search for Adamic language proceeds in the service of praising God for the uniqueness he bestowed upon us when creating Adam, the first speaker. Rather, Adam's offspring searches for the ideal formal language in the service of science, making them unique in nature in virtue of being masters of that very same nature. Russell (1914/2009), a follower of Frege across La Manche, argued that our everyday natural language is not sufficiently abstract and adequate to convey the true structure of reality which is to be discovered and neatly reported by science. Only carefully selected properties of natural language were deemed worthy enough to be included in formal language. Such

property was compositionality, which conveys that the meaning of a whole sentence is determined by the meaning of its constitutive parts. Thus, compositionality became central formal device for the ideal language, and Frege's shoes were later filled in by proponents of TGG, LOT, and traditional symbolic cognitive science who took compositionality to be essential feature allowing for mapping between language and thought. Frege also stipulated that the structure of ideal formal language should be in tune with the structure of thought as Ockham did in the 14th century, i.e., such language for Frege was indeed ideal *qua* transparent medium of thought/semantic content. As Tyler Burge describes, "...Frege was primarily interested in eternal structure of thought, of cognitive contents, not in conventional linguistic meaning" (1979: 213). In later works, Frege (1918/1977) confined such conceived thoughts to the Platonic Third Realm, different from physical world and inner world of individual speakers: they were thought to be immaterial immutable, and graspable by any human being, thereby ensuring the intersubjectivity of the meaning, as opposed to fluctuating conventional meaning.

Thus, there is a close relationship between thoughts and meaning in Frege's philosophy of language. His famous (1892/1952) sense and reference distinction (ger. *Sinn und Bedeutung*) grounded two-level semantics by showing how proper names and sentences could have the same object of reference but differ in terms of their cognitive significance. Take a rational (sic!) subject *S*, who has conflicting beliefs about the same object of reference, e.g., *S* may believe that Scipio Africanus Minor ordered the complete destruction of Carthage but highly doubt that Scipio Aemilianus had anything to do with waging a war against Carthage, since she is unaware that both "Scipio Aemilianus" and "Scipio Africanus Minor" refer to the person. The truth value of a sentence "Scipio Aemilianus ordered the destruction of Carthage" and "Scipio Africanus Minor ordered the destruction of Carthage" is a function of truth-values of its constituents—as per principle of compositionality, obviously, and in this case both sentences are true. Proper names are co-referential and pick out the same person, whereas the denotation od the predicate is function that takes "Carthage" as argument or object. However, differences in cognitive significance of these sentences arise, most notably in propositional attitudes expressing beliefs, doubts, desires, etc. For *S*, therefore, the two sentences are incompatible in virtue of being claims about different persons, and for each of them, *S* associates specific modes of presentations, or in Russell's (1905) lingo, descriptions. The sense of proper names amounts to descriptions. Thus, at the level of reference, the sentences are true (even trivially true), but at the level of sense, they are not identical due to *S* doubting the contingent fact that Scipio Aemilianus and Scipio Africanus Minor are the same person. *S*, therefore, entertains two different thoughts given that the sense of a sentence is thought (ger. *Gedanke*). What is Frege's solution for reference fixing in tricky cases like the one I previously described? He simply relies on the rationality of individuals: no rational subject could, at the same time, hold two contradictory beliefs. Frege's sense vs. reference distinction and emphasis on the first-person perspective gave rise to internalism and descriptive theory of reference in philosophy of language and paves the ground for another distinction, namely extension vs. intension.

Internalism is a position in the philosophy of language which states that meaning is solely dependent on internal factors, i.e., by what is happening inside the speaker's head. Her associations, clusters of descriptions as in Searle (1958), representations, modes of presentations, intentions, basically whatever we stipulate to be within one's skull and mind, can exercise influence on the individuation of semantic content and, in turn, picking out reference—and behold the link between the descriptive theory of reference and internalism. Internalism can also be taken to align with rationalism inasmuch one assumes that the speaker's rational grasp accounts for the linguistic meaning. Truth-conditional semantics was overwhelmingly internalist at its inception given that the subdiscipline can be construed as providing us with models of minimalist idealizations of the semantic properties of language without taking into account any external factors embedded in communicative practice (Harris 2017: 174-175). However, this is quite different from historical rationalism that was intertwined with the innateness of particular beliefs.

The peak of the ideal language project within the truth-conditional semantics came with Richard Montague (1970a/1974, 1970b/1974, 1970c/1974) who strived to formalize natural language *en général* rather than take the apt excerpts from it that are stipulated to be suitable for logical analysis. This subproject was called *Montague gramma*r and outlined the sharp distinction between linguistics and philosophy of language in terms of goals and methodology. While Chomsky was interested in syntax and envisaged TGG as primarily syntax-oriented theory, Montague admittedly did not see any merit in syntax except as a preliminary for semantics and looked down on TGG: "One could also object to existing syntactical efforts by Chomsky and his associates on grounds of adequacy, mathematical precision, and elegance" (1970b/1974: 223, fn. 2). Thus, the formal ambitions of linguistics were nowhere near the established pedigree of truth-conditional semantics in philosophy of language. Montague endorsed both Frege's famous distinction between sense and reference in the form of intension and extension, as well as principle of compositionality to derive his formal semantics, which was subsumed under the motto *syntax is an algebra, semantics is an algebra, and meaning is a homomorphism between them* (Janssen & Zimmermann 2021). His approach was extensional, which essentially means that Montague considered intensions as functions from possible worlds to extensions. The ramification of this choice was reluctance to go along the path of intensional logic, since an extensional approach was considered, along the Fregean lines, to be more rigorous.

Unlike Frege, however, he managed to suggest mathematical tools for the logical analysis of the aspects of natural language that Frege bypassed, i.e., imperatives and questions as typical non-declarative sentences for which truth conditions are not applicable but rather fulfillment conditions (Montague 1970c/1974: 248, fn. 3). According to Harris (2017), this marked the end of truth-conditional semantics *qua* models of minimalist idealizations of language fragments and turn towards empirical investigation of meaning. Partee (1980) sketched two parallel views of semantics, one being mathematical, the other psychological. Either one takes semantics to be representative of semantic competence and can be investigated from the processing perspective, or one considers semantics as isolated from any psychological or biological details, and Partee (1980: 4-5) found Montague grammar to be inherently incompatible with two quite basic insights from psychology and

cognitive science, namely that our brains are finite (as opposed to Montague's intensions of sentences amount to functions from possible words to truth values) and we are endowed with knowledge of our language and its constitutive meanings or concepts, viz., semantic competence (whereas Montague's intensions of words are functions from possible worlds to extensions).[10]

Despite Montague's keen insights and virtuosity, the challenges to the ideal language project came rushing from three fronts, one being from the rise of ordinary language philosophy, second was dug out with the inception of externalism and causal theory of reference, and third, the most painful, came from one of their own who debunked two dogmas of empiricism but also rooted out the core tenets of the project. Empiricism in the 20th century was spread across these three fronts, often varying in intensity and type of commitment. Let me start from the third front, the most intense and explicit in its commitments to historical empiricism. Philosophers and scientists gathered in the Vienna circle, called logical positivists, were sympathetic to both Frege's program of providing logical foundations for the mathematics and ideal language project and impressed with British empiricism, which they found to be in the accordance with the burgeoning scientific progress of the 20th century. These two quite different commitments were embedded in their verifactionist principle, which stated that the sentence is meaningful if and only if it is justifiable by empirical methods, which boils down to sense experience (see e.g., Carnap's (1928/1967) *Aufbau* project). Thus, only scientific sentences have meaning. Mathematics and logic are, on the other hand, vacuously true in the sense that they do not offer any meaningful i.e., cognitively significant, information about the way the world is. Simply put, mathematics and logic are analytic truths. The ramification of the endorsement of such philosophy of language was that pretty much anything can be found to be meaningless unless based in sensory experience—even the verifactionist principle *per se* (see e.g., Hempel 1950). Empiricist scientific rigor and the quite literal application of the ideal language project within logical positivism proved to be self-defeating. In the 1950s, logical positivism was shuttered by one of its most loyal *protégés*, Willard Van Orman Quine, who challenged both verificationist principle and the view that mathematics and logic are vacuously true, i.e., each of the two dogmas of empiricism.

Quine (1951) discarded the idea that it makes any sense to verify or confirm individual scientific sentences because these are always intertwined within a specific scientific theory. Furthermore, his minutiose analysis of different notions of analyticity showed that there is little reason to claim that mathematics and logic are vacuously true and independent from truth-conditions for physical world—no sentence is immune to revision, even those that were traditionally taken to be analytic truths. *A fortiori*, the traditional distinction between analytic truths and synthetic truths, i.e., verifiable with

[10] In Subotić (2017) I showed on a small fragment of Latin how Montague's formal analysis of that-clauses can be applied, especially for contexts involving subjunctive rather than usual indicative. However, such a formal analysis Montague is wildly psychologically implausible given the psycholinguistic studies of human parsing of that-clauses. This suggests that the conviction that natural and formal languages can be treated on a par thanks to mathematical tools can only take you so far. I turned to connectionism precisely because of the promise of ever-increasing psychological plausibility of natural language processing.

respect to the physical world, is doomed and of no use in a consistent empiricist view. In Sect. **3.2.**, by relying on Buckner (2023), I present the third dogma of empiricism, which is concerned with establishing theoretical support for connectionist account of linguistic competence by describing language acquisition through domain-general mechanisms, of which Quine would probably be sympathetic to at least some extent. Let me explain the reasons for my confidence.

In the 1960s, Quine proceeded to design his empiricist system that encompassed epistemology, philosophy of science, philosophy of language, and philosophy of logic. The starting point was that all knowledge, including scientific knowledge, starts with stimulation of sensory nerves, and the very entry into corpus of knowledge begins with observation sentences. Being one of the rare philosophers of language who discusses the process of the acquisition of linguistic meaning, Quine left the ivory tower of analytic philosophy in the search for grounding his views on meaning psychology, most notably behaviorism. Thus, he needed to envisage a coherent, psychologically backed up story of how children learn observation sentences, and thus get cognitively inaugurated into linguistic community. In *Word and Object* (1960/2013), Quine gradually set the scene: the child is thought to have disposition to assent or dissent in response to either direct or indirect stimuli, and the corrective role of the linguistic community, most often parents, help actualize this disposition by punishing or rewarding specific behavior. Mastering sentence production, and language usage generally, hinges on the child's ability to acquire relevant sets of syntactic rules and semantic representations—to which the linguistic community conforms—in relation to the set of directly or indirectly experienced situations.[11] Thus, for Quine, the very notion of meaning is not confined to the ideal language project in any way but is rather a behavioral phenomenon which deserves empiricist treatment. Similarly, the new connectionist dogma that will be introduced in Sect. **3.2.** sees language as an emergent usage-based phenomenon. Additionally, as I will show a bit earlier, in the Sect. **2.1.**, Chomsky's criticism of behaviorism *à la* Skinner, which Quine wholeheartedly incorporated in *Word and Object*, as well as Chomsky vs. Quine dispute about language acquisition, offer glimpse into what will be crammed into the

[11] The acquisition of linguistic meaning is described within Quine's (1960/2013: 29) more developed and well-known though experiment about the indeterminacy of translation, in which a field-linguist investigates an indigenous language by keeping a diary on speakers' behavior and strings of sounds associated with such a behavior, and when a rabbit crosses the path of one of the speakers, he utters *Gavagai*. Our field-linguist, in Quine's *mise en scène*, cannot be sure whether the speaker refers to the rabbit, parts of the rabbit, etc. The reason why she has to adhere to eliciting either assent or dissent of speakers in the presence of stimulus lies in the fact that only via behavior she can correlate reactions and gesticulation with the purported linguistic meaning. In this way, our field linguist can uncover stimulus synonymity without imposing sets of syntactic rules and semantic representations of her own language to the indigenous one that she is investigating (Quine 1960/2013: 52-53). Stimulus analyticity, then, amounts to communally endorsed observational sentences for which the speakers have consensus, i.e., show almost universal assent (Quine 1960/2013: 66). The *Gavagai* thought experiment prompted Quine to believe that linguistics must be married to behaviorism, while maybe in psychology one could have an alternative to behaviorism that could be equally scientific and rigorous (see his 1987 paper).

rationalist/nativist argumentative line *contra* accounts that deny the innateness of language faculty like connectionism does.

However, for the time being, let me return to the two left fronts. Both were intertwisted with the third, albeit the intensity and the type of commitment to empiricism were much lower and much looser. Quine' rejection of the analytic/synthetic distinction provoked the answer from philosophers of language who had already believed that both Frege's and logical positivists' programs are inadequate *qua* methodology and orientation in philosophy. Thus, Grice & Strawson (1956) defended the distinction on the grounds that it has its function in ordinary language. In Locke's manner, these authors thought that natural language should be in focus, specifically the communicative practice of ordinary speakers, since this could shed light on philosophical problems (see also e.g., Wittgenstein 1953, Austin 1962). The ordinary language philosophy originated in G. E. Moore's common-sense approach to philosophy, and as Burge succinctly puts it: "[t]he tradition deriving from Frege took science, logic, or mathematics as the source for linguistic and philosophical investigation, whereas the tradition deriving from Moore took ordinary practice as the touchstone for linguistic and philosophical judgment" (1992: 12). Some of the philosophers inclined to this orientation combined ordinary language philosophy with behaviorism, which resulted in logical or philosophical behaviorism which stated that terms referring to mental states gain their meaning in virtue of being customary used to convey the disposition to a particular behavior (e.g., Ryle 1949) as opposed to Descartes's conviction that there is *res cogitans* on the inside, i.e., the entity with mental states being only its modes. Inasmuch as this strand was committed to behaviorism, it was inclined to empiricism as well, albeit this was more of an implicit commitment.

The gentle pull towards the usage of ordinary language and Quine's arguments *vis-à-vis* linguistic behavior that could be evaluated by fellow speakers contributed to creating a favorable moment for a more radical attack on the internalism and descriptive theory of reference. Philosophers who looked down on the ordinary language admitted that ordinary language philosophers were right in pointing out that the ideal language project has severed ties between linguistic meaning and reality too abruptly. There are many intricacies and anomalies of natural language that are quite informative and interesting for developing better formal tools for describing them. Thus, these authors aimed to reform natural language by ascribing to it an equal level of importance and value as formal languages (Burge 1992: 15). As opposed to Montague's extensional approach, their approach was intensional. Kripke (1972), who devised intensional, modal logic as means to account for modalities present in the natural language, offered a series of examples showing that clusters of descriptions are neither sufficient nor necessary for a speaker or a community of speakers to fix reference. His alternative, the causal theory of reference, stipulated that the reference is fixed by the initial act of baptism, and then via causal chains the meaning spreads across generations, and in each generation, the speaker relies on others for continual transmission and fixing of reference. During this process, reference may end up distorted or change, and allowing this within a theory of reference shows the significant maturity of philosophy of language to deal the with language *qua* both synchronic and diachronic phenomenon. Putnam (1975) further emphasized the cognitive division of labor,

i.e., deferring to experts for reference fixing, and with it social and environmental factors as decisive for determining the meaning of singular terms, most notably natural kind terms. His motto—*Cut the pie anyway you want, meaning just ain't in head*—is usually taken as constitutive for externalism, the position that emphasizes the importance of external factors for determining the meaning instead of mental repertoire of speakers that figured in internalism. Semantic content, according to externalists, is individuated in terms of one's relation to community and physical environment (Burge 1992: 25). Externalists may be characterized like empiricists on a long stick, especially if one adjoins the additional commitment to the nature of language acquisition device by connecting it to community and physical environment. However, as in the case of internalism, the position was silent about the origins and different from historical empiricism.

What happened with philosophy of language after the 1970s? Furthermore, whither the empiricist vs. rationalist debate in the 20th century? As Burge (1992) nicely describes in his largely first-hand overview of the history of philosophy of language and mind, the demise of the philosophy of language *qua prima philosophia* began with its increased specialization which suggested that the discipline exhausted its promise in successful dealing with traditional philosophical problems despite the empirical boost in the form of behaviorism. Thus, philosophers of language turned to fine-grained analysis of usage and shifted towards specific pragmatic phenomena, or indexicals, which overwhelmingly remained in the Fregean framework, stuck between being a peculiar anomaly of natural language and the challenge for the lurking ideal language project that remained a pipe dream (although *cf*. Sect. **4.3.**, where I mention one of the few accounts of indexicals that were not orthodox). Virtually nothing could have stopped the renaissance of nativism embodied in Chomsky's TGG, which I will be tackling in Sect. **2.1.**, given that behaviorism was also ostracized. On the other hand, with the genesis of cognitive science in 1978, the attention naturally shifted to the philosophy of mind, and the positions of internalism and externalism gained new connotation, which I will be presenting in Sect. **2.2.** that is coming down the pike. This shift and intrigues surrounding it when it comes to linguistic competence will be the main topic of the next Ch. As for the rationalist vs. empiricist debate, it seems that what was a historical *a priori* debate, with relatively clear frontlines, became a chaotic fire at will in the 20th-century philosophy of language, with scattered theoretical commitments across the polarities internalism/descriptive theory of reference and externalism/causal theory of references.[12] The *confusio linguarum* progressed into intra and interdisciplinary confusion about linguistic competence *qua* theoretical entity in years to come.

[12] Truth be told, this is more of a continuum given that philosophers, as is tradition, lumped together some of the positions to form hybrid theories of reference. However, for simplicity sake, allow me to depict the situation like this.

2. RATIONALIST AND EMPIRICIST ASSUMPTIONS IN COGNITIVE SCIENCE AND LINGUISTICS

We know too little about mental structures to advance dogmatic claims.

—Noam Chomsky (1980: 49)

2.1. The Renaissance of Rationalism, pt. 1: Transformational-Generative Grammar

The historical debate between rationalists and empiricists was construed around a *priori* beliefs, such as mathematical and logical truths, or anything related to God. The psychological question (*What are the origins of these beliefs exactly?*) was ancillary to the epistemological question (*What are the means for justifying these beliefs?*), which was, in turn, answered with "they are implanted in our minds by a benevolent God [who had] better things to do than spend time splicing beliefs into psyches at the appropriate experiential moment..." (Cowie 1999: 27). Thus, linguistic capability as means for expressing those beliefs was in the middle of the debate during the early modern period: the rationalist camp considered it as the most precious gift from benevolent God which makes us unique, while the empiricist camp sought to provide a sensory-based image of this capability that incorporates mundane communicative situations. The 20th-century philosophy of language divided linguistic capability into two inter-related acts, namely act of referring and act of understanding the meaning. Different accounts of meaning and reference fluctuated around the primacy of natural language over formal language or *vice versa*, extensional over intensional approaches, as well as around the argumentation about whether internal or external factors fix the reference.

Rationalist echoes were traceable in internalism, given that both positions were searching for ideal language that would mirror the structure of thought and avoid imperfections of natural language, albeit with a significant difference that lies in the fact that internalism was silent about the innateness. Empiricist echoes were traceable in externalism, given that these philosophers found merit in the otherwise notorious natural language: its imperfections are less of an issue if one takes into account communal or environmental context. However, as discussions in the philosophy of language got entangled, the implicit commitments became messier. Quine, at the same time an offspring of logical positivists and Fregean extensional approach, endorsed behaviorism as psychological theory in which he embedded his philosophy of language, which, in turn, led him to empiricism. Hence, it was believed that some meek form of empiricism was the mainstream in the 1970s given the efforts of Quine and the proponents of causal theory of reference, so Noam Chomsky's treatment of language seemed like a counter-revolution or even renaissance to philosophers given his allegiance to rationalism and nativism (Hook 1969: x). In this Ch., I turn to the intertwined intellectual histories of linguistics and cognitive science. Here, I am more inclined to tracing chronology and linear order per Sect. in comparison to the previous Ch. Curiously, this strategy will allow me to pinpoint almost always neglected corners of cognitive science and cognitive neuroscience, which could

radically shift our perspective on both rationalism and empiricism professed within these scientific disciplines in the 20th century.

Be it as it may, from the 1950s, with its peak in the 1970s, nativism concerning linguistic competence was back on the table for the first time after three centuries. On the one hand, Chomsky considered himself to be doing work that has not much to do with philosophy of language except for putting kibosh on behaviorism (in 1959), which directly affected most of Quine's arguments as well (see Chomsky 1969). His area of specialization was syntax, whereas philosophers of language were oriented towards semantics. On the other hand, Chomsky cared a great deal to provide a pedigree for his own work in linguistics, which, curiously, made him turn to the dignified past of philosophy, specifically Cartesian legacy, rather than linguistics. As I was showing in Sect. **1.1.** & **1.2.**, this *modus operandi* was far from being flawless given that one can trace Cartesian ideas much earlier and in a more complex format than Chomsky cared to present. In what follows, I tackle upon linguistic roots of TGG, as well as broader theoretical, and most notably philosophical context in which the development of various versions of TGG unfolded during the 20th century.

American linguistics was largely independent from and unaware of the philosophy of language, but equally smitten with behaviorism. Chomsky's advisor, Zellig Harris, was one of the key figures of structuralism. The structuralist like Harris held that the main job of linguistics is descriptive: by employing an axiomatic discovery procedure, linguists should discover compact, simple constituents of all possible structures within a corpus that contains functional units at phonetic, morphological, lexical, and syntactic level, i.e., phonemes, morphemes, words, phrases, and sentences (Seuren 1998: 214). Interestingly, however, structuralists also believed that theoretical terms such as "words", "phrases", etc. should be defined with respect to their observable features such as sequence of sounds, which puts them in the behaviorist camp. Thus, in their view, linguists should only catalogize regularities in a corpus and steer clear from assuming any corresponding but unobservable mental catalogue to natural language.[13]

Harris's axiomatic discovery procedure, pioneered in his 1951 book, was a trailblazer for Chomsky's formal approach. The procedure takes rules, envisaged as axioms, for predicting the constituents of sentences in a corpus from which deductive system was being built, and this system then leads to the first set of theorems about relations between sentence

[13] Quine was most probably acquainted with Harris' *Methods in Structural Linguistics*, since Harris here hinted at an issue for which Quine's stimulus analyticity could provide a solution: "It is possible for different linguists, working on the same material, to set up different phonemic and morphemic elements [...] The only result of such differences will be a correlative difference in the final statement as to what the utterances consist of"(1951: 2). Thus, Harris basically claimed that functional analysis proceeds in a manner that individual linguists see fit. However, as we have seen in fn. 11, when doing actual fieldwork, i.e., when the corpus is not known in advance, but rather under construction, linguist's choice is underdetermined by the available evidence. Thus, stimulus analyticity is needed for fixing at least some parts of the corpus which would serve as reference points. Interestingly, Chomsky's criticism of behaviorism generally did not coincide with the criticism of his advisor who was also fond of behaviorism. Rather, as I will be showing in the next paragraphs, Chomsky's criticism of Harris was either misconstrued or pointlessly exaggerated, but never included any mention of behaviorism.

constituents, as well as to the second set of theorems which indicate the type of structures present in a corpus (Harris 1951: 372). Harris is, in fact, preoccupied with the mathematical description of surface structure, and still does not refer in any way to deep structure, that Chomsky introduced. Be it as it may, Seuren notes that "here we have, *in nucleo*, the concept of generative grammar [...] Note that we do not have any notion of transformational grammar yet [...] But we will not have to wait long..." (1998: 228).

Harris's main contribution that influenced Chomsky is two-fold. First, he shifts perspective from individual sentences to the totality of sentences in corpus. Second, he looks for the best generative rule system. Unfortunately, Harris's system was highly impractical since it presupposed that linguists would first analyze sounds to make a phoneme inventory, then proceed to morphemes, and continue all the way to sentential level. Some six years later, nonetheless, Zellig Harris discovered transformation from "horizontal", i.e., surface structures as analyzed in his 1951 book, to "vertical" generative operations (Seuren 1998: 238). Vertical operations allowed for ordering of the previously analyzed structures. As Seuren (1998: 239, fn. 20, 241, fn. 22) argues, Chomsky largely ignored Harris's turn towards transformations or presented their views as pitted against each other because he considered Harris to be instrumentalist rather than realist about structures, whereas Harris considered their views as offering complementary tools for advancing linguistics. Seuren also stresses Chomsky's tendency to present his work in linguistics as unprecedent, solitary and entrenched "in a much more distant and dignified past [,] for which he did not develop interest until the early 1960s, and then only in so far as it could be used to 'legitimize' his own points of view" (1998: 250, fn. 26).

In 1957, Chomsky published *Syntactic Structures*, where he gave a more precise account of transformational rules and integrated it into generative grammar, thereby paving the way for TGG as the new post-structural paradigm in linguistics.[14] This core is not going to change despite other alterations of the TGG framework in years to come. The core amounts to three ordered sets of rules: formation rules that generate deep underlying syntactic structure, transformation rules that generate surface syntactic structure, as well as phonological, morphophonemic, and purely morphological rules that generate phonological and morphophonological representations. Somewhat counterintuitively,

[14] *Caveat*: Throughout the dissertation I use the term "paradigm" in the non-Kuhnian sense. I am more inclined to consider TGG, connectionism, and traditional symbolic science as research programs in the Lakatosian sense. The first reason behind this terminological decision is that TGG is far from being Kuhn's (1962) "normal science" despite the efforts of Chomsky and his followers to make it look like mainstream in linguistics. The framework of TGG has been frequently changed and tweaked (*ipso facto* it is rather progressive in Lakatos's (1978) terms) and alternatives are being envisaged from the very beginning, such as generative semantics and cognitive linguistics, which I mention (and wholeheartedly endorse) in Sect. **2.2.** and Ch. **4**, respectively. Connectionism, being inherently future-biased (see Sect. **3.1.**), would be at best seen as immature science in Kuhnian framework, despite being actively used as alternative "paradigm" to symbolic cognitive science for the past 40 years (*at least*, do see Sect. **2.3.**). Besides, neither symbolic cognitive scientists nor connectionists thought these were incommensurable approaches with respect to their terminology, since the former camp considered the models of the latter camp as mere implementations of symbolic architectures albeit with a biological flavor. Lakatos's account of active comparison and evaluation of research programs seems like a better fit here.

though, Chomsky seldom delved into descriptive and grammatical analysis of English corpus as Harris did. One reason for this is that, unlike typical structuralists, he put forward the view of grammar *qua* device that can generate infinite number of sentences, i.e., even those outside any known corpus, through the finite set of sentences which are evaluated as well-formed by speakers of a given language. Thus, what matters is formal procedure and the core of TGG rather than mere cataloging of natural language utterances. The other reason probably has something to do with Chomsky's penchant for metatheoretical questions in linguistics, which made him more of a metalinguist than linguist (Seuren 1998: 252-255). For instance, with the advent of TGG, he introduced moderate realism in linguistics that soon became mainstream thanks to the genesis of cognitive science. Moderate realism in linguistics assumed that linguistic theories should approximate cognitive machinery in humans, i.e., any hypothesis about linguistic competence and language processing must be formulated with respect to speakers' cognitive architecture.

Specific assumptions about human cognitive machinery embodied in the idea of innate linguistic competence led Chomsky to crusade against behaviorism. The world saw yet another book in 1957 that marked the recent history, namely Skinner's *Verbal Behavior*, that had the mission to reinforce behaviorism *qua* dominant experimental and theoretical framework in psychology. The crux of Chomsky's (1959) argumentation in the negative review of Skinner's book is that language, being creative and productive capacity given that we are able to produce sentence about things we never heard about or seen, must be determined to some extent by an internalized grammar which constrains the space of all possible sentences. Skinner (1957: 107-108), on the other hand, held that children learn language via reinforcement contingencies that are controlled by verbal community: children are either punished for not replying to specific stimuli or rewarded, and in this way, their control of stimuli is actively being sharpened by behavior modification. The goal is to become an effective speaker as well as listener, given that children become initiated into verbal community by listening to sounds coming out of their caretakers' mouth. The older they get; the verbal context becomes more important for developing verbal competence since some features of stimuli are more salient than others and influence generalization in children (see Skinner 1957: 331-334). As I will be discussing in Sect. **2.3.** and Ch. **3**, the functioning of ANNs could be reminiscent of some Skinnerian points, although they should not be equated with behaviorism *tout court*. Chomsky was not trying to refute Skinner premise by premise but *en masse*:

> "The fact that all normal children acquire essentially comparable grammars of great complexity with remarkable rapidity suggests that human beings are somehow specially designed to do this" (1959: 57).

Virtually the same conclusion was reached after his brief exchange with Quine: Thus began the new era of TGG as per commitment to moderate realism and anti-behaviorism: once the formal syntactic structures were spelled out, the corresponding cognitive machinery had to be found, and we already know what we are looking for—something that makes us *special*.

Moderate realism in linguistics is always accompanied by further commitments such as commitment related to the nature of data that are relevant for hypothesis testing.[15] As opposed to data in terms of physical sounds that Zellig Harris favored, Chomsky was focused on the idealized data about grammaticality, or well-formedness of sentences of one's mother tongue. He explained that "Linguistic theory is concerned primarily with an ideal speaker-listener (…) who (…) is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance" (1965: 3-4). Thus, in his next book, *Aspects of the Theory of Syntax*, Chomsky made explicit his distinction between competence and performance that I will be eager to discard in Ch. **4**: the ideal knowledge of grammar is linguistic competence *sensu stricto*, whereas the actual use of the language is a matter of less relevant linguistic performance. Linguistic theory should abstract away from performance issues altogether and dedicate efforts to unraveling linguistic competence, i.e., innate rules that allow native speakers to differentiate between well- and ill-formed sentences or grammatical and ungrammatical ones. There are no grades of grammaticality, as if rules bring with them exclusive disjunction. The rules are usually subsumed under the much more popular term "universal grammar".

The *Poverty of stimulus argument* was the main argument for universal grammar, presented in a rudimentary form in Chomsky (1965: 58) but sharpened throughout the 1970s and, curiously, still is the MVP of TGG.[16] Although there have been numerous formulations of the argument in the literature so far, allow me to present my own that strives to avoid technicalities:

(1) A child in virtue of being a teeny-tiny scientist who acquires a language via hypothesis testing does so thanks to its tacit knowledge of grammar or linguistic competence.
(2) Given that a child is faced with low-quality and scanty primary linguistic data, it seems unlikely that the child could become competent in syntactic structures as efficiently as it actually does.
(3) Therefore, the child who acquires a language must know a lot about language in advance (*from 1 & 2*).

[15] *NB*: Moderate realism in linguistics aligned well with representational realism in the traditional symbolic cognitive science thereby constituting a unified front against any alternative that would dare to cast doubt on realism *qua* metatheoretical position or commitment. As I will be showing in Sect. **2.3.**, this is one of the reasons why shallow connectionism could not throw the baby out with bathwater, i.e., argue against rule-based in favor of the pattern-based cognition *and* endorse instrumentalism about representations.

[16] I owe the reader an important *caveat* here: due to limited space, I will not be discussing the Poverty of stimulus argument at great length with respect to its empirical inadequacy that has been piling over the years since this would merit a paper or dissertation on its own. The empirical evidence against this argument was cited and thoroughly discussed with respect to its philosophical and cognitive significance in, e.g., Elman et al. (1996), Cowie (1999), Clark & Lappin (2011), and Dabrowska (2015). All these authors primarily attack nativism on the grounds that it is not supported by the poverty of stimulus argument given that neither premises nor conclusion were validated. My attack on the argument is to be found in Sect. **4.2.**

(4) Having tacit knowledge of grammar, being endowed with linguistic competence, or knowing a lot about language in advance is compatible with nativism, the view that some domains of knowledge are innate (*from 1-3*).
(5) Therefore, a child has an innate linguistic competence (*from 1-5*).

Thus, premises (1), (2), and (3) identify a gap between competence and performance, and Chomsky thought this gap would be sufficient to establish that no general learning mechanism put forward by an empiricist could account for this gap, hence (5) would naturally follow. However, this is a most uncommitting form of the Poverty of stimulus argument given that it is silent about the exact mechanism that links nativism to having tacit knowledge of grammar, i.e., it is unclear in what sense some domains are innate and why one label linguistic competence as such. Post-1965 formulations of the argument were under a quite surprising influence of the philosopher from the different camp, namely Putnam (1967).

Putnam purported to show that one could accept the Poverty of stimulus argument without committing to nativism as in (4) since there is nothing special in language *qua* domain of knowledge unless Chomsky first offers solid reasons for (4). Domain-general multipurpose learning mechanisms could govern language acquisition in children that had been proven in other domains: any organism will rely on recursive mechanisms to make sense of its experience with environment in form of the patterns and strive to choose the most simple and informative patterns. Thus, for instance, linguistic universals are to be preferred over complex grammars that lack them, and it is unclear why such universals would constitute evidence for domain-specific grammar. Nonetheless, what is crystal clear is that the argument amounts to *a posteriori* premises and conclusion, as shown above, which further means that these are *de facto* empirical hypotheses, and indeed, Chomsky offers no empirical data to back up (4) (see Cowie 1999: Ch. 8). Note that no amount of dignified philosophical legacy could back up this argument despite Chomsky's ambition professed in *Cartesian Linguistics* (1966), because the historical schism between rationalism and empiricism was aprioristic. Rather, one could say that Chomsky's aims, as well as argumentative strategy, are theoretical and have remained so, whereas it is up to his followers to scrap the empirical evidence. From the 1970s, the efforts of TGG turned to finding universals, universal constraints, or principles that would be constitutive of the universal grammar conceived as a domain-specific module governing language learning and shared among all (and only) humans. The Adamic language thus got its cognitive extension in the form of linguistic competence *à la* Chomsky.

The most promising line of inquiry that could have made of universal grammar an empirical hypothesis par excellence, was presented in Chomsky's *Principles and Parameters* (1980). There, he argued that principles are fixed, whereas parameters can be switched on or off depending on the particular natural language. *Ab initio*, all parameters are switched off, and this amounts to the initial state of universal grammar. As child begins his inauguration into community, primary linguistic data affect only some parameters, like, for instance, *pluralia tantum* parameter for Serbian language, whereas such parameter is in the off position for the speakers of Vietnamese, Japanese, or Korean. However, this program was found to be implausible on evolutionary grounds given that it would be highly unlikely

that grammar of such complexity could have emerged as early as archaeological evidence suggests (Terzian 2021: 3). In one of the latest versions of universal grammar, embedded in the so-called *Minimalist program*, Chomsky is explicit about his general programmatic hypothesis about looking for that one abstract language in the world—akin to the Adamic language—that could further be parametrized to obtain variations that we see as multifarious natural languages. But this is merely a mirage. In his words:

"[a] narrow conjecture is that there is no such variation: beyond [phonetic form] options and lexical arbitrariness (which I henceforth ignore), variation is limited to nonsubstantive parts of the lexicon and general properties of lexical items. If so, there is only *one* computational system and *one* lexicon, apart from this limited kind of variety" (1995/2015: 155, my emphasis).

In the most recent book, with an emblematic title *Why Only Us* (2016), Chomsky turned to evolutionary arguments in favor of the uniqueness of human linguistic capabilities instead of producing manifestos every ten years.[17]

Nonetheless, Pieter Seuren, an eminent linguist with a disdain toward Chomsky's ideological manifestos that are scarcely entrenched in empirical evidence and actively immunized against falsification by his followers on the grounds that counterexamples or negative evidence are peripheral and contaminated by performance factors, gives the following assessment of TGG tradition:

"The ideological urge to provide backing for whatever unifying principle that was being considered at any given time often has led to a selective presentation of data and far-fetched explanations for unsupportive facts that could not be ignored...There is, moreover, a distinct tendency among Chomskyans to suggest that they have a special, privileged access to the mysteries of language, a hot line to heaven, so to speak..." (1998: 283-284).

As I will be arguing in Sect. **3.1.** and **3.2.**, not much has changed in the 21st century among rationalists in cognitive science or AI research who ground their position in the Chomskyan nativism when it comes to computational models of NLP. The rhetoric of linguistic competence being unique domain which cannot conform to any other hypothesis about cognitive architecture except symbolic reinforces the line of argumentation that *any* computational model designed in non-symbolic paradigm, regardless of its successful performance on NLP tasks, is doomed from the very start. Alas, for some, the hot line to heaven seems to be always on hold.

[17] I return to Chomsky's recent contributions in Ch. **4**, where I advance several arguments against them as well as against the core of TGG—namely, the idea of rule-governed linguistic competence and the competence vs. performance distinction. A separate line of criticism can be put forward against his account of the evolution of linguistic competence as in Chomsky & Berwick (2016), albeit it is out of this dissertation's scope, but see e.g., Martins & Boeckx (2019) for a solid line of criticism.

2.2. The Renaissance of Rationalism, pt. 2: Representational Theory of Mind and Traditional Symbolic Cognitive Science

With the demise of behaviorism and the growing myopia of philosophers of language, linguistics was regarded almost as identical to TGG. However, TGG was focused on syntax and its relation to lexicon because Chomsky thought that any discussion of word and sentence meaning should be quarantined to philosophy so that a sharp demarcation line could be drawn between scientific and non-scientific analysis of language. Moreover, a stronger argument was being advanced by Chomsky that syntax should be regarded as cognitively autonomous. Nonetheless, Generative Semantics was introduced by, *inter alia*, George Lakoff, Jerry Fodor, Paul Postal, James McCawley, and Pieter Seuren. They were all interested in deep structures and convinced that the level of abstractness present in Chomsky (1965) was insufficient to account for subtle semantic differences on the surface level, such as quantifier scope or odd and seemingly meaningless sentences (Seuren 1998: 493). The shallow roots of the novel account of deep structures were removed as early as the 1970s: some of the generative semanticists preferred Chomsky as a friend rather than truth (*pace*, ancient philosophers), Chomsky and his followers were either way winning the battle for funds and graduate students, the rest of generative semanticists were relatively unfamiliar with the developments in logic which left them unable to make use of formalisms, and, finally, Montague's virtuosity that led to the inauguration of formal semantics, or Montague grammar (Sect. **1.3.**), stopped generative semanticists in their tracks altogether (Seuren 1998: Subsect. 7.3.1 & 7.3.2). Put this way, the story sounds relatively simple and not at all dramatic. However, history is rarely simple and without drama, even in the case of scientific history.

The so-called *Linguistics Wars* over deep structures ensued in the late 1960s and continued throughout 1970s and left a scar on linguistics as Randy Allen Harris recounts in his recent history of linguistics (2021). The generative semanticists who stayed under Chomsky's wing believed that semantics was still inferior to syntax in the sense that meaning is derivative from syntactic structure. They soon became adherents of Interpretative Semantics. Jerry Fodor was one of them. The renegade generative semanticists, led by George Lakoff, were more radical and considered semantics equally important as and independent from syntax. As Randy Allen Harris neatly sums it up: "At one end, generative semanticists argued that language was one big schmoosh, with no place at all for borders (…) At the other end, Chomsky's camp, the interpretative semanticists, seemed to be boundary fetishists, redrawing their borders daily; one day, a piece of data was syntactic, the next day morphological; one day it was semantic (…) Each saw other side as perverse, and each opened its guns on perverted" (2021: 10).

Essentially, the dispute targeted competence vs. performance distinction. Chomskyans and Chomsky held competence in high regard due to its syntactic purity and cut off any semantic anomaly to performance section. George Lakoff and his comrades regarded this division as *ad hoc* and theoretically sterile, while the nature of language is more patchy, messy and knows no artificial divisions into competence and performance. Take, for instance, the famous example from Chomsky's *Syntactic Structures* (1957):

"Colorless green ideas sleep furiously". The first camp saw this as an instance of syntactically well-formed, i.e., *grammatical* sentence, albeit semantically meaningless and unacceptable to speakers. For them, this proved that syntactic and semantic processing occur independently, and the job of TGG is confined to determining grammatical sentences not acceptable ones. The allowed gloss over this treatment of such sentences amounted to post-hoc semantic interpretation that took syntactic structure as input and meaning or lack thereof as output. The generative semanticists, however, believed that such sentences show to what extent Chomsky and interpretative semanticists miss something crucial about language: the whole gamut of fine-grained levels of grammaticality and acceptability influencing each other given that semantics pierces through syntax. Generative semanticists lacked cognitive backup story which TGG aficionados had in the form of endorsed nativism, but that changed in the 1980s when cognitive linguistics emerged on the shoulders of cognitive science.

The Linguistic Wars preceded the *Brain Wars* but, curiously, revolved around the same issue and argumentative lines, as Joe Pater argues on his blog titled "Brain Wars". The Brain Wars ensued after the conception of cognitive science and only deepened the dispute over deep structure between Chomsky and Lakoff with a clear convergence towards rationalist or empiricist position *vis-à-vis* origins of linguistic capacity. Soon enough, the Brain Wars took a general form of representing a bloody feud between the two competing research programs reflecting either rationalist or empiricist allegiance, namely symbolic and connectionist approaches to human cognition. This boiled down to the *AI Wars* between GOFAI and ANNs as preferred methodologies for modelling human cognition, which last to this day, thus making AI Wars continuous with both Linguistic and Brain Wars. History has a way of repeating itself, even in a few years' time span. The very topic of the rest of this and the next Ch. will be the mesh of the Brain and AI Wars.

One of the veterans of the Linguistics Wars, Jerry Fodor continued his battle for syntax-first-semantics-second *credo* coupled with nativism within cognitive science, a burgeoning, promising new field that appeared in the late 1970s. The backbone of this new scientific discipline was philosophy of mind, the new philosophical game in town that was seen as worthy heir of philosophy of language, along with linguistics (Chomskyan), psychology (now without the heavy burden of behaviorism), computer science/AI research (alive and kicking since Dartmouth conference in 1956)[18], anthropology, and neuroscience.

[18] Since the main topic of the dissertation is linguistic competence, I focus on the intertwined histories of philosophy, linguistics, and cognitive science, which means that I omit details from the history of computer science/AI research, albeit with a guilty conscience. The decision to skim over this piece of intellectual history was made due to the scope of the thesis, but I strive to mention some important figures and events either in footnotes or in relation to cognitive science. In this spirit, it is worth mentioning Dartmouth workshop which gathered all the pioneers of computer sciences and early AI, from John McCarthy, Marvin Minsky, Oliver Selfridge, Claude Shannon, Allen Newell, Herbert Simon. During the couple of weeks, they discussed their ideas in an open-minded manner, including whether to focus on deductive or inductive reasoning in machines, digital or analog computation, expert systems or systems that learn. After Dartmouth workshop, in September, another event took place at MIT, namely symposium of the Special Interest Group in Information Theory, in which some of the Dartmouth experts also participated. Miller (2003: 142-143), being

Functionalism was put forward in philosophy of mind with the hope that this position will be able to account for mental states in tune with psychology and computer science and avoid the pitfalls of behaviorism.

Functionalists believed that a mental state M_n is inextricably linked to input, output, and a network of M_1, M_2, M_3, *etc.*. In other words, mental states are defined through their function in a wider system made of all the other mental states, stimuli which prompted it, and behavior. Some of the versions of functionalism were analytic, i.e., conceptual and disinterested in ramifications of the position for mind and brain sciences, while the others, such as machine functionalism and psychofunctionalism, sought to clearand maintain the metaphysical and conceptual ground for scientific inquiry (see Levin 2023 for a review). The common denominator for all versions was the conviction that functionalism abstracts away from physical realization of mental states, i.e., introduces multiple realizability of the mental. Creatures different from humans may be described as having mental states since these can equally be realized in biological bodies and, say, silicon, because identification of these states proceed with respect to their function. This made functionalism relevant for AI research and, therefore, an ally to the Computational theory of mind (CTM). CTM is the view that the mind is akin to a computational system *du jour*—this may be a Turing machine, a digital computer, or whatever machine is popular at the moment *and* performs computations. CTM was at first intertwined with machine functionalism as proposed by Putnam (1967), who emphasized functional isomorphism between states of stochastic Turing machine and mental states. Operations of the machine are specified by explicit instructions such as: If in t_1, state S_1 receives input I_1, then in t_2 there is probability p that S_1 would go to S_2 and result in output O_2. Functional isomorphism, in fact, maps machine states such as S_1 and S_2 onto mental states. *Ergo*, human mind is akin to computational mind, such as stochastic Turing machine.

However, Block & Fodor (1972) put forward an important criticism of machine functionalism that will figure prominently in the rest of this dissertation only with a different target. This criticism was based on the notions of productivity and systematicity of *thought*: human beings are capable of entertaining infinite number of thoughts based on finite number of elements constituting such thoughts and having a capacity for entertaining one complex thought entails entertaining other simpler or similar thoughts. Machine states, on the other hand, are finite, unstructured and holistic, and thus lack the sensitivity of structure that would allow for systematicity. In other words, this version of CTM that aligns with functionalism was deemed not rich enough. A subtype of CTM, namely Representational theory of mind (RTM), was envisaged as a theoretical skeleton for the idea that mind is akin to digital computer, specifically its software.[19] One of the main proponents

---

one of the speakers, witnessed the planting of the seeds of cognitive science during the second day of the symposium when, among others, Newell & Simon presented their work in AI, David Hebb his neurological theory of cell assemblies, and Noam Chomsky showed how information theory may be used for syntax analysis.

[19] *NB*: RTM as a subtype of CTM can also be defined as richer, symbolic CTM. Connectionism, which I will present in the next Sect., can also be considered as a representational and computational theory, albeit with an utterly divergent connotation of both what it means to compute or represent something.

and pioneers of RTM was none other than Jerry Fodor. According to this theory, mental states and processes are intentional—they are *about something* and, therefore, have semantic content that can be evaluated with respect to its truth-conditions. Such semantic content was dubbed "mental representation". Mental representations amount to the objects of propositional attitudes like beliefs, hopes, etc. Thus, vehicles of semantic content carry the content and, at least in Fodor's (1975) *Weltanschauung*, language is a medium for mental representations, much like Ockham envisaged the relation between language and thought. Thought processes amount to the causal sequences of the tokening of mental representations (Fodor 1987). Since thoughts occur in language, RTM, in fact, postulates mental states that are computational in virtue of instantiating mental representations *qua* symbols of the so-called Language of Thought. Fodor's Language of Thought hypothesis (LOT) was the first of its kind after six long centuries and influenced by the interplay between linguistics, philosophy of mind, psychology, and AI research.

Note, however, that in Block & Fodor's (1972) paper, their argumentation did not refer to language-bound properties such as productivity and systematicity but took them as a tacit assumption. Only a couple of years later, Fodor started doing philosophy of mind from the interpretative semanticist's point of view by starting with the productivity which is, as per Chomskyan framework, essential for language due to its intrinsic relation to linguistic creativity (i.e., the possibility to make up sentences that were never written or uttered given the infinitely many combinations one can make from finite number of structures and rules). What made RTM superior to Putnam's machine state functionalism coupled with CTM was exactly its power to account for both productivity and systematicity thanks to its secret ingredient LOT. Thus, RTM postulated simple and complex mental representations—simple ones may be only concepts such as TIBERIUS, AGRARIAN REFORM, ROMAN REPUBLIC, whereas complex ones are structure-sensitive like TIBERIUS PROPOSED AGRARIAN REFORM IN THE ROMAN REPUBLIC.[20] With Frege's blessing, Fodor applies compositional semantics to mental representations—complex representations are functions of the structure and content of its simpler constituents. With these clarifications of Fodor's RTM in mind (besides all the mental representations), let us see how RTM handles productivity and systematicity. A finite set of simple thoughts can easily be combined to produce infinitely many complex thoughts, maybe even those that were never entertained so far (I will optimistically offer the following: PITY TIBERIUS DID NOT PROPOSE THE AGRARIAN REFORM IN DAGESTAN). Furthermore, it is easy to account for systematic relations between simple and complex thoughts alike due to structure-sensitivity (a follow up on the earlier example—PITY DAGESTAN HASN'T GOT SOMEONE LIKE TIBERIUS TO PROPOSE THE AGRARIAN REFORM). Or, in Fodor's manner of speech:

"OK, so here's the argument: Linguistic capacities are systematic, and that's because sentences have constituent structure. But cognitive capacities are systematic too, and that must be

[20] I use caps lock to convey the concept or thought in LOT and to distinguish them from a regular natural language word or sentence.

because of thoughts. But if thoughts have constituent structure, then LOT is true. So I win and Aunty loses. Goody!" (1987: 151)

As a true disciple of Chomsky, Fodor (1975) was a believer in nativism—all simple lexical concepts, which are building blocks of mental representations, should be construed as innate in the LOT framework, whereas complex ones may be acquired (although *cf*. his 2008 book). In other words, Fodor only transferred Chomsky's methodology and convictions to the domain of semantics, even in the case of assumptions about the origins of cognitive mechanisms behind semantics.[21] However, although Chomsky looked for rationalist legacy in early modern philosophy, it is much more difficult to locate Fodor's philosophical legacy. This is mostly because RTM muddied the waters of different philosophical positions with respect to mind and language. That is, one could believe that the way to individuate semantic content is through either internal or external factors, i.e., factors pertaining to one's own intrinsic properties, or those pertaining to one's environment and community. Tyler Burge (1979, 1992) introduced the distinction individualism vs. anti-individualism to pinpoint this difference by extending Putnam's externalism about linguistic meaning to propositional attitudes. Thus, it is quite uncontroversial to be internalist and individualist in this regard—Fodor emphasized structure-sensitivity of semantic content, which is intrinsic to one's mind given LOT. However, in his 1987 book, Fodor contended that semantic content has to be causally dependent on the outer world as well, thereby cleaving closer to externalism and anti-individualism rather than internalism about semantic content.

RTM and LOT were seen as coextensive to the practice of cognitive science from the 1970s, and hence abductively true: these were the best tools we had for shedding light on the inner cognitive machinery that finally came under scrutiny after human mind, i.e., the black box of behaviorism, had been opened up. All the key components of traditional symbolic science were already present in Fodor's work—taking the mind-as-digital-computer metaphor at face value, endorsing LOT and RTM, professing allegiance to nativism, considering TGG as the mainspring of the successful treatment of higher cognitive processes that goes to show our unique place in the nature.

As George Miller, a psychologist who stood at the forefront of the new interdisciplinary scientific field, recounts, the first step towards unified mind science were made at three universities in the USA—Harvard, Carnegie Mellon, and the University of California in San Diego—during the 1970s. However, Miller sees 1978 as the official

[21] This association, although almost never explicitly stated, guides also the work of contemporary philosophers of linguistics who have a penchant for LOT. Thus, for instance, Gabe Dupre (2020) combines LOT and TGG to show how internal language (I-language in Chomsky's terminology, which amounts to linguistic competence) is much like LOT in the sense that it deviates from external language (E-language in Chomsky's terminology, which amounts to linguistic performance), and instead of going for the gap between language and thought in the case of ungrammatical but acceptable linguistic expressions, one would be better off with showing how such expressions generate deficient structures and ill-formed thoughts. Thus for Dupre, the language of thought in LOT is natural language as understood and examined in TGG. I go in completely opposite direction in Ch. **4**, and argue in favor of decoupling thought from language by mounting a case for connectionism.

birthdate of cognitive science when Alfred P. Sloan Foundation provided a number of universities with grants for establishing programs and centers that would be dedicated to cognitive science. Cognitive science is constituted by six fields (thus, often represented as the multidisciplinary hexagon, see Fig. 4 below), namely philosophy, linguistics, psychology, computer science/AI, neuroscience, and anthropology, as well as their mutual cross-pollination, which was in the ether and unofficially practiced from the 1950s (Miller 2003: 142).[22] At its core, cognitive science treated mind as software, brain as hardware, functionalism as a means to make sense of this relation between mind/brain and digital computer, and language as the pinnacle of human rational thought whose mechanisms are best conceived as innate and isolated from the rest of basic cognitive machinery, body, or environment—which was really a stark contrast to cultural evolution that anthropology insists on for all the other cultural tools. The first interdisciplinary field that emerged as a direct result from the Sloan Foundation grant was cognitive neuroscience at the Cornell Medical School, which relates theoretical foundations of cognitive science with neurobiological evidence and computational modelling as preferred methodology. As I will be showing in Ch. 3, contemporary connectionism has further blurred the line between cognitive science, neuroscience, and cognitive neuroscience, often levitating between the three fields.

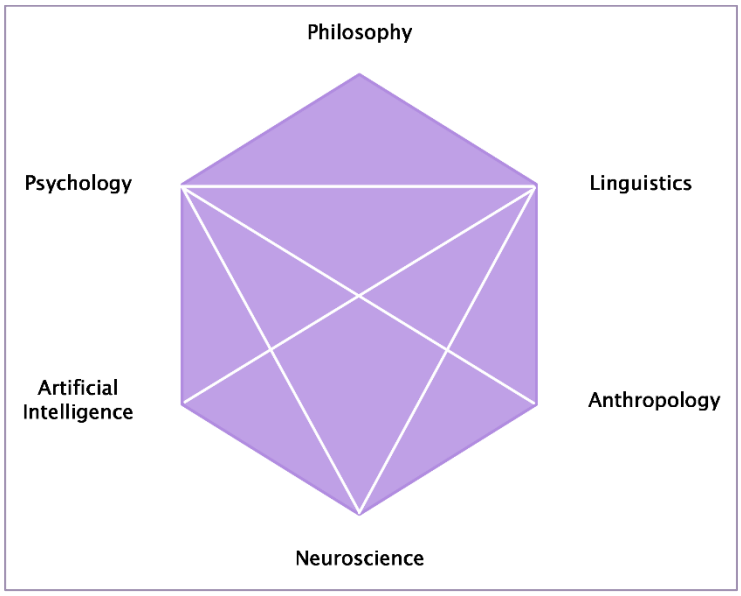


Fig. 4 The cognitive science hexagon

The early cognitive science was later dubbed "traditional symbolic cognitive science" due to its theoretical and methodological commitments, which constitute a

[22] The links between fields give rise to various subdisciplines. In the 1970s, there were five of them, as represented in Fig. 4, but it is safe to say that nowadays all 15 links have been forged, at least on the micro-level, i.e., level of the publications of individual scientists, philosophers, and AI engineers. However, as Núñez et al. (2019) point out, although the links have been forged, cognitive science did not manage to emerge as a cohesive, mature scientific field with integrated theories and unifying methodology. Moreover, their bibliometric and institutional indicators suggest that interdisciplinarity on this macro-level has never been established, but rather cognitive psychology basically devours all the other fields and curricula are surprisingly myopic in this regard as well.

particular strand (or epoch) in cognitive science. As Rogers & McClelland (2014) summarize, cognitive science may be defined as the effort to answer three questions:

> **(i)** *What kinds of processes support the complex behavior of intelligent systems*;
>
> **(ii)** *What kinds of mental representations do such processes operate on*;
>
> **(iii)** *What is the origin of such processes and representations, i.e., are they innate or learnable through experience*?

Traditional symbolic cognitive science, which both incorporated and was influenced by TGG, provided the following answers to these three questions:

> **(i$_S$)** *Cognitive processes are like digital computer programs – they resemble ordered lists of explicit or implicit rules; and they are sequential, which means that each process follows domain-specific rules and that each process waits for its predecessor to end so that the appropriate output could be computed*;
>
> **(ii$_S$)** *Representations are discrete and symbolic. They have syntax and compositional semantics, which means that structurally molecular representations have syntactic constituents that are themselves either structurally molecular or atomic and that the semantic content of a molecular representation is a function of the semantic contents of its syntactic constituents;*
>
> **(iii$_S$)** *Mechanisms underlying cognitive processes are best understood as innate since the number of possible ordered lists of rules is virtually unbounded, so the initial constraints must be prespecified rather than learned*.

All three claims **(i$_S$)**-**(iii$_S$)** can be seen as generalized assumptions of TGG to the cognition as a whole—recall that Chomsky also saw universal grammar as the innate rule-governed syntactic device. And as I was showing in this Subsect., Fodor virtually mapped TGG onto his treatment of semantics. Also, note that **(ii$_S$)** is basically the essence of Fodor's LOT, allowing thus symbolic cognitive architecture to account for productivity and systematicity of thought because the thought resembles language. Moreover, symbolic cognitive science is confined to the personal level of psychological explanation of cognitive phenomena, and, thus, committed to the realism about propositional attitudes. This means that beliefs, desires, etc., as conceived and used in common sense or folk psychology, i.e., to describe and understand the behavior of others, have their scientific implementation within symbolic cognitive science *qua* posits of folk psychology.

This further implies that the semantic content is to be found at the personal level in symbolic cognitive science: content has conceptual and syntactic structure as presumed by LOT and humans are attributed with intentionality on the basis of taking propositional attitudes seriously within symbolic paradigm. Now, recall, semantic content can be either individualist/internalist or anti-individualist/externalist (world-involving, if you like), so the labeling of the proponents of symbolic cognitive science is trickier than the labeling of generativists who are self-proclaimed rationalists. Furthermore, being committed to either internalism or externalism with respect to content is not the same thing as being either internalist or externalist about particular vehicles of content, which are subpersonal events

or properties relevant for psychological explanation of cognitive phenomena (see Dennet 1991). Being a vehicle internalist amounts to turning to the internal physiological or neurological events, while vehicle externalist would rather draw on the external relations between an organism and its body or environment (Hurley 1998: 3). Radical vehicle externalist would regard either body or environmental factors as being constitutive of mental states/cognitive processes thereby extending their realization outside of the skull and biological boundaries of humans (e.g., as in Clark & Chalmers 1998). Symbolic cognitive science is rationalist in spirit inasmuch one sticks to claim **(iii)**, i.e., that cognitive mechanisms are innate and domain-specific, and the disembodied view of cognitive processes, i.e., internalism about vehicles of content, but its philosophical underpinnings with respect to semantic content significantly muddy the waters *vis-à-vis* nature of cognitive processes since one can coherently adopt internalism about vehicles but externalism about content.

Finally, let me introduce computational modeling of natural language processing within traditional symbolic cognitive science. Both symbolic and connectionist cognitive science, which will be the topic of the next Sect. (and the rest of the dissertation for that matter), have the ambition to offer a computational modeling methodology and a hypothesis about the cognitive architecture within the human mind. This can be reformulated as the search for adequate computational architecture that will explain and simulate our cognitive architecture faithfully enough. Finding the right cognitive architecture would, in turn, help us identify the essential properties of linguistic competence that makes us unique. Both strands consider their methodology and the view of cognitive architecture as the only true account of human cognition, so the stakes are quite high. So far, I have been describing what constitutes symbolic cognitive architecture, and this straightforwardly maps onto the modeling methodology, which is implicitly suggested in claim **(i)**. i.e., that cognitive processes resemble digital computer's program. Traditional symbolic cognitive science was seduced by Good-Old-Fashioned-Artificial-Intelligence (GOFAI) that roamed around the corners of computer science and cybernetics departments during the 1970s as the only game in town. The GOFAI was seen as the mimicry of human deductive capabilities and the realization of wildest dreams of all the philosophers and logicians who preferred formal language and logic over natural language and mundane inductive reasoning. Unsurprisingly, the GOFAI modeled geometry, spatial reasoning, algebra and deduction so well that it was thought to capture the very quintessence of human intelligence along the rationalist lines because it made use of hardwired, manually specified rules to manipulate symbols (Newell & Simon 1961). This was eerily similar to the rationalist account of innate knowledge which give rise to God's blessings such as language and mathematical and logical truths (see Haugeland 1985: Ch. 1)

Theoretical and technological advances in GOFAI in the 1970s allowed for computational models of natural language processing that basically were extensions of the

earlier models concerned with formal languages, logical and spatial reasoning.[23] Winograd's (1972) model SHRDLU was able to converse in natural language about rearranging blocks on the table since it was fed in advance with a detailed script of instructions, albeit this constituted too much of a narrow and artificially created domain to be considered a genuine contribution to the cognitive analysis of NLP. Soon enough, parsing models that aimed to entrench TGG in computational framework overflowed cognitive science and psychology. These models purported to simulate the succession of transformations from surface to deep structures in the tree-like hierarchical structure but ran into difficulties when contrasted with data from psycholinguistic studies of human syntax processing (Christiansen & Chater 2008: 480 and references therein). Further development of expert systems, i.e., symbolic models that were designed for problem-solving based on larger but narrow body of knowledge, most notably expert knowledge, and explicit instructions in the forms of rules, allowed for a full-blooded generativist treatment of language processing. Dissected from all other general and/or sensory processes, parsing procedure in a typical symbolic model maps input onto partial syntactic structures per rules of the specified grammar and gives categorical output thereby vindicating the idea that linguistic competence operates as fine-tuned selection device for grammatical as opposed to ungrammatical sentences (Gibson 1998, cf. Christiansen & Chater 2008: 479). Nonetheless, even these models were all too brittle and—despite the noble (and historically long) idea of idealized linguistic competence that God, natural selection, or innate mechanisms endowed us with—all to implausible with respect to actual, human NLP, full of noisy performance factors. The time was ripe for another revolution, or better yet—a reform.

2.3. The New Wave of Empiricism: Connectionist Cognitive Science

Connectionism, or Parallel Distributed Processing (PDP) as initially dubbed, was the underdog of cognitive science from the very start. Associated with virtually all diametrically opposite positions, theoretical commitments, and methodology, connectionism was a direct archnemesis to everything traditional symbolic cognitive science stood for. Connectionism *qua* computational modeling approach relied on ANNs composed of units classified into at least three layers, which learned from data in input to produce the output with a higher probability thanks to the stored processing signal in hidden units, and, therefore, to better account for data that were not part of the training set. The training of an ANN proceeds via specific learning algorithm which impacts connections between units in a given layer, and each unit has activation threshold. The strength or intensity of these connections is called weight (Braddon-Mitchell & Jackson 2007: 220). By adjusting weights, ANNs learn to produce the adequate output with respect to input. The ANNs described here were classical shallow feed forward ANNs (FNNs) in which every

[23] These advances came at the expense of connectionist modeling given that all the funds were redistributed to GOFAI thereby resulting in the first AI Winter according to the canonical view of the history of connectionism. In the next Sect., I provide more details of this historical period in the AI research and – spoiler alert – cast doubt on the canonical view.

unit in a layer is connected to all units in the next layer, but units within a single layer do not interact with each other. This means that there was no cycle of information flow in FNN.

Most histories and historical overviews see connectionism as the new wave of the 1980s (e.g., see Buckner & Garson 2019, Berkeley 2019), which rested on the pioneering work of psychologist Frank Rosenblatt who introduced perceptrons—first multilayered FNNs for pattern recognition—that were destroyed by philippics in Minsky & Papert (1969) in favor of GOFAI. This course of action brought the first AI Winter. However, this image of connectionism is too simplified and inaccurate. My guess was that this canonical image served connectionist well to represent the research program as revolutionary in the 1980s in cognitive science. However, this would be mere rhetoric. ANNs *qua* engineering feat had been around from the 1950s, i.e., almost as long as GOFAI, the first such model being Oliver Selfridge's *Pandemonium* (1958).[24] This model was designed to account for image constancy in a biologically plausible way by postulating independent feature or letter detectors that are parallelly connected, amusingly named data demons, computation demons, and cognitive demons. Current research on letter perception is based on Selfridge's key idea that images of letters are detectable due to their component features that constitute perception patterns, and the models are even referred to as "pandemonium-like models" (Grainger, Rey, & Dufau 2008).

From 1958 to 1962, Frank Rosenblatt developed perceptrons on the *Mark I Perceptron Machine* at Cornell thanks to the funds of the Office for Naval Research because such pattern recognition software was seen as potentially useful for advancing geospatial intelligence on the eve of Cuban Missile Crisis (see O'Connor 2022). After Rosenblatt's untimely death, a number of AI engineers continued to refine pattern classifications through the conception of machines that "learn to learn" such as Nilsson (1965) and Ivakhnenko & Lapa (1965). Moreover, the ground was made fertile for the birth of connectionist cognitive science already in the 1960s, given that Rosenblatt's perceptrons were so influential that cognitive psychologists wrote textbooks from the perspective of pattern-based cognition (see Neisser 1967). It is safe to notice that connectionism started as a mainstream option a decade before the genesis of traditional symbolic cognitive science, but also crashed and burned quickly. Interestingly, even though adversarial relationship between Rosenblatt and AI pioneer Marvin Minsky is always mentioned in the literature, it is seldom known that Minsky closely collaborated with Selfridge on *Pandemonium* and the motor-driven potentiometer inside Mark I, which encoded weights of photocells or units, was built seven years earlier by none other than Minsky (Anderson & Rosenfeld 2000: 304), therefore sharpening the

[24] The impressive pioneering work of McCulloch & Pitts (1943) is also often used as a starting point for connectionism since the dynamic duo argued that neurons can be axiomatized as performing logical operations, and thus recreated as artificial neurons. They went on to apply their results in neuroscience by analyzing frog vision (see interview with Jeremy Lettvin in Anderson & Rosenfeld 2000: 1-21). However, strictly speaking, this is not a really accurate picture: despite being focused on neurons, thereby being suggestive of ANNs, McCulloch & Pitts were inspired by Leibniz and in this rationalist spirit wanted to try out the idea that a neuron can be formal, symbolic device, which makes them more aligned with traditional symbolic science, or at best with implementational connectionism, i.e., connectionism that implements symbolic cognitive architecture or LOT.

image of connectionism as the institutional mainstream in both AI and mind sciences during the 1950s and 1960s.

Anyhow, by the time Minsky co-authored a book with Papert (1969) on the shortcomings of perceptrons, such as the inability of three-layered perceptrons to compute exclusive disjunction if the units produce local representations, The Defense Advanced Research Projects Agency (DARPA) habitually contacted academic researchers who could provide them with technology that could buttress efforts of the US military to prevail in the arms race with the USSR. Thus, Olazaran (1996) argues that Minsky & Papert's criticism of perceptron was fueled with self-interest and struggle to obtain funds for GOFAI rather than waging the AI Wars for the sake of scientific truth, especially because major funding of ANN-based AI halted in the 1970s. Nonetheless, after Lighthill Report (1973), which brought about the first AI Winter in the United Kingdom due to its takeaway message that most AI researchers just did not live up to their end of bargain since most models were mere toy models with no real-world application, DARPA cut funding for GOFAI as well. However, given that GOFAI found its application in traditional symbolic cognitive science, researchers continued to rake in the dough as opposed to ANN aficionados, now scattered around departments for neuroscience, where they continued to work on learning algorithms for ANNs. As one of the indications that work on ANNs had not been deterred by the lack of funding in the 1970s was the development of the most influential learning algorithm, namely backpropagation. Backpropagation is a key component in training ANNs using gradient-based optimization algorithms like gradient descent. Gradient descent works by iteratively adjusting the model's parameters in the direction of the steepest descent of the loss function because the goal is to find the set of weights and biases that result in the best performance of the model. The function here should be understood as non-linear.[25] Backpropagation computes the gradient of the function so that the network's parameters can be adjusted to minimize errors that arise in the output after processing data. That is, the error is backpropagated through the layers of the feed-forward ANN. A form of this algorithm was first used in 1974 for training ANNs (Werbos 1974). Werbos (1982) was also the first to describe a successful application of backpropagation for efficiently training ANNs, although this did not turn many heads like Feldman & Ballard (1982) did. As Robert Hecht-Nielsen, one of the San Diego-based connectionists, recalls in his interview (Anderson & Rosenfeld 2000: 298-299), around 1982 he pitched ANNs to DARPA as being valuable tools for radar development. Only then did the so-called connectionist revolution in cognitive science begin.

[25] The output of layers of units within ANN is computed by taking a weighted sum of unit inputs and passing that sum through an activation function, which can be either linear or non-linear. If the activation function is linear, then ANN is severely limited in its ability to capture any complex patterns. Nonetheless, the advent of backpropagation allowed for avoiding this linearity by enabling the ANN to approximate nonlinear functions because backpropagation computes gradients throughout the layers of non-linear activation functions. Sigmoid activation function was the most used in connectionist modeling—it maps the input values between 1 and 0 thereby facilitating binary classification. This type of non-linear function suffers from vanishing gradient: it squashes the input to a small range, which, in turn, results in gradients becoming extremely small, so the updates to the units' weights become negligible and hinder the rate of learning.

The ambition of connectionist modelers in cognitive science was to limit as much as possible the manual specification of learning parameters and to let the ANN learn by itself. A parameter is a variable that is automatically optimized during the training process and a hyperparameter is a parameter that must be set before the training process, usually directly by a modeler (see **Tab. 1** for examples). Hyperparameters are used for controlling the learning procedure, i.e., the process of mapping independent to dependent variables. However, modelers should carefully choose the number of parameters and hyperparameters for fear that ANNs may be sullied by overfitting: when ANNs have large number of parameters relative to the training data that is available to them, they tend to overfit, i.e., to produce outputs that are too similar to data which hinders ANNs predictive powers.

Tab. 1 A list of parameters and hyperparameters for connectionist models

Parameters	*Hyperparameters*
Weights	Number of weights, choice of activation function (e.g., sigmoid)
	Choice of optimization algorithm (e.g., gradient descent)
Inductive biases	Number of hidden layers
	Learning rate, number of iterations or epochs during training

The mission of minimizing parameters and rules made connectionism *qua* hypothesis about cognitive architecture dedicated to proving that nativism is obsolete. Thus, for instance, PARSNIP was a model implementing a FNN that learned grammatical structures from exposure to sentences present in corpus, and the modelers were quite explicit in their intention to make PARSNIP as much as possible free from encoded syntactic structures and rules governing the prediction of the next grammatical category (Hanson & Kegl 1987: 107).[26] However, the main reason for giving up on PARSNIP was its inadequacy regarding the plausibility of sentence processing: since FNNs had no means to account for dynamics of processing, i.e., to incorporate time, PARSNIP did not accurately capture NLP. In this regard, the connectionist strand was a rightful heir to British empiricism, concerned with biological and psychological plausibility of cognitive processing. In other words, the aim was to pinpoint sufficient and necessary physiological, neurological, and psychological constraints on processing in order to support functional isomorphism between a model and human mind/brain. As stated in the PDP Bible:

"Though the appeal of PDP models is definitely enhanced by their physiological plausibility and neural inspiration (...); [w]e are, after all, cognitive scientists and PDP models appeal to us for psychological and computational reasons. They hold out the hope of offering computationally

[26] Truth be told, PARSNIP was trained via supervised learning which comprised annotated data with respect to grammatical categories. The modelers did not really have anything better at hand, although their enthusiasm for avoiding nativism shows the connectionist spirit *par excellence*.

sufficient and psychologically accurate *mechanistic* accounts of the phenomena of human cognition (...); and they have radically altered the way we think about the time-course of processing, the nature of representation, and the mechanism of learning" (Rumelhart & McClelland 1986: 11, my emphasis).

As the quote suggests, mechanistic framework was a good match for connectionism: the models' inner mechanisms of functioning, indicative of model's capacities, should be regarded as a core component of *explanans* of human capacities. Or, in other words, mechanisms are realizers of functional isomorphism at a different level of explanatory grain than was the case in the traditional symbolic cognitive science. Let me clarify this. The only *prima facie* similarity lay in the fact that connectionism was also committed to representationalism like symbolic cognitive science, albeit connectionist representations came in a completely different format. Representations in connectionism are *vector representations* showing patterns of weighted connections among units, and instead of being localized, they are *parallelly distributed* throughout the network. Vector representations encode input data into numerical format where every vector dimension encodes a specific feature or aspect of data. Contrary to symbolic representations, connectionist vector representations are not structure-sensitive, but rely on finding patterns in data that should correspond to and/or predict patterns in world. Moreover, connectionist models are concerned with real world models and real-world data, as opposed to toy models of symbolic cognitive science, as explicitly stated both in the PDP Bible (see e.g., Ch. **4**) and by the pioneers of ANNs such as Terrence Sejnowski (in Anderson & Rosenfeld 2000: 331).[27] The link between a model and target system is already present in the design of connectionist models given the simulated environment for the training and testing ANNs. The environment is represented via time-varying stochastic function over the vector space of input data (Rumelhart & McClelland 1986: 53-54).

This makes connectionism in line with externalism and anti-individualism of semantic content. It is important to note that semantic content in connectionism is not on the personal level as in the traditional symbolic cognitive science, but rather on subpersonal or subsymbolic level (Smolensky 1988). Thus, connectionism is not committed to realism about propositional attitudes nor is concerned with macro-cognition, i.e., intentional mental states. To quote from the PDP Bible: "In general, from the PDP point of view, the objects referred to in macrostructural models of cognitive processing are seen as approximate descriptions of emergent properties of the microstructure" (Rumelhart & McClelland 1986: 12). The main issue with connectionist ambitions to offer the competitive hypothesis of cognitive architecture was the lack of structural representations and psychological explanations at the personal level. For this reason, most criticism, most eminent being Fodor

[27] This is also closely intertwined with the issue of brittleness of symbolic models: once they are unable to match instruction with the task, the model fails to produce any output at all, unlike humans, who do not simply abandon task when they are unfamiliar with the instructions. On the other hand, connectionist models are blessed with graceful degradation (Rumelhart & McClelland 1986: 29), which means that task performance comes in degrees: the units with misleading features may activate wrong output, but through backpropagation, the model will continue to look for the state of equilibrium, instead of reacting fatally to errors. Hence, in this sense, connectionist models are more aligned with the flexibility of human behavior.

& Pylyshyn (1988), reckoned that connectionism could be seen, at best, as the exotic implementation of LOT, or else it is not much different from the notorious behaviorism.[28] In other words, the brain may well be a connectionist machine, but thought is systematic and productive, and the only way to account for these two essential features is through LOT. Connectionist models could be trained to exhibit systematicity with respect to their behavior, but they are not nomologically systematic, thereby making hypothesized cognitive architecture inapt to be considered as a viable and autonomous account of human cognition (Braddon-Mitchell & Jackson 2007: 228). Fodor & Pylyshyn also add compositionality *à la* Frege to this list of essential features as well as inferential coherence. Thus, for them a cognitive architecture must be a combo of "syntactically driven machine whose state transitions satisfy semantical criteria of coherence" (1988: 30).

The duo did not successfully demarcate all these features. For instance, productivity and systematicity differ only in latter being intrinsically linked to the ability to produce and/or understand sentences, while the former assumes that we are *in principle* able to produce an infinite number of sentences despite finite cognitive apparatus. Compositionality is introduced to underpin systematicity by pointing out that systematic sentences are not like that by chance but owing to semantic contribution of each and every sentence constitutent. Productivity does not entail systematic compositionality, even though it is closely intertwined with compositionality and systematicity. But in any case, it is possible to have partially productive and partially non-systematic domains (see Baroni 2019: 3).[29] Chemero (2011) considers Fodor & Pylyshyn's criticism an instance of the Hegelian argument, i.e., a conceptual criticism with no empirical support that authoritatively asserts that *p*, *p* being here the claim that connectionism will fail *qua* hypothesis about cognitive architecture (or become assimilated into the symbolic strand). Their authoritative stance opened the gate for many ecumenical solutions over the years, especially in the domain of language processing (e.g., Steedman 1999, Pater 2019), thereby normalizing the implementational status of connectionism as opposed to its coveted autonomy. However, ironically, Fodor & Pylyshyn offer no empirical evidence that human thought is systematic, but draw this from the implicitly assumed LOT. Connectionists are still allowed to deny either that the thought is systematic, productive, or compositional, or that natural language, as used in everyday communciation, lacks any of the features given that these features were singled out based on considerations pertaining to formal language

[28] What is truly impressive is that these two lines of criticism remained the same (*verbatim*!) from the 1980s (Fodor & Pylyshyn 1988) up to 2021 (Childers, Hvorecky, & Majer 2021). The former I tackle in the Subsect. **The New Wave**, the latter I comment in the Subsect. **A Strawperson Empiricist and Impartial Rationalist Enter a Bar**. This is mostly impressive because connectionist models are completely different in the 21st century in comparison to the 1980s, and for this reason could be conceived as future-biased research programs (see Sect. **3.1.**). Alas, their criticism can easily be seen as past-biased and ideological rather than constructive and to the point.

[29] Teaser: in Sect. **3.3.** however, I argue that maybe a way out of the rabbit hole is to regard post-connectionist models of NLP as showing us how (some parts of) language can be productive without necessarily being systematic in the strict sense *à la* Fodor & Pylyshyn (1988). I encircle this line of argumentation with the general idea that the connectionist paradigm teaches us a valuable lesson about decoupling language from thought.

or formal accounts of natural language.[30] Chemero maintains that neither the cognitive science of the 1980s nor of the 2000s words is mature science but rather an immature and underdeveloped one, with "no universally accepted paradigm", hence "background assumptions that structure the research of one faction are optional to those of other factions" (2012: 15).

Finally, what about the vehicles of semantic content if this content resides on the subpersonal level? At first, internalism regarding vehicles of content was a more natural ally to connectionism, given its goal to be the alternative to symbolic models based on the biological flavor, which made connectionist models more flexible. But from the dawn of the 21st century, a turn towards externalist vehicles has been steered given the pleas for embodiment and boosting of the good old biological flavor with situated environment (see Sect. **3.1.** & **4.3.**). Let me now summarize the differences between connectionism and symbolic cognitive science by enlisting connectionist answers to the three questions singled out by Rogers & McClelland (2014):

> **(ic)** *Cognitive processes are like analog computer programs because they are modeled in such a way that the primary aim is to find the most highly associated output corresponding to an arbitrary input within the ANN. Weights of connections between input units and output units are adjusted until the statistical properties of input units are recapitulated among the environmental events. The detection of statistical patterns is performed by hidden units that are not directly connected to the environment as other units are.*
>
> **(iic)** *Representations are parallelly or neurologically distributed within a neural network. By giving a complete, formal and precise account of microlevel, or subsymbolic level–where states of units' activation correspond to patterns of statistical and neural activity–it is possible to simultaneously obtain approximately true generalizations at the macrolevel, or symbolic level.*
>
> **(iiic)** *Knowledge in an ANN is learnable from experience with data and environmental factors encountered during training phase. A plethora of learning procedures are available in connectionist research: backpropagation or error correction, Hebbian learning, etc.*[31]

Connectionism is dedicated to vindicating empiricism about the mind, which is professed in **(iiic)** but also in its commitment to externalist content (and vehicles). The PDP

[30] In Subotić (2018) and partially in Subotić & Milojević (2021), I have covered the intricate details of the 1980s clash between symbolic cognitive science and connectionism, providing an overview of the heated argument exchange between key figures regarding the sentence processing, and commenting the systematicity issue, which spans across myriads of publications in the last three decades, often amounting to the highly technical and narrowly-conceived disagreements (see e.g., Smolensky 1987, Fodor & McLaughlin 1990, Fodor 1997, Matthews 1997, also a whole edited volume by Paco & Calvo 2014). I will not rehearse any of this here because I want to return to the issue from the perspective of the latest DL models for NLP—which offer reasons for optimism.

[31] And soon enough, DL algorithms. The progress of connectionist modeling is directly correlated to the development of more efficient algorithms, an increase in computational power so that bigger amounts of data could be part of the training set, and, finally, architectural novelties. Check Sect. **3.1.** for details in this regard.

Bible is also outspoken in its allegiance to empiricism as opposed to nativism and rule-governed cognition constitutive of the traditional symbolic cognitive science:

> "The approach that we take (...) is completely different. (...) [W]e do not assume that the goal of learning is the formulation of explicit rules. Rather, we assume it is the *acquisition* of connection strengths which allow a network of simple units to act as though it knew the rules" (Rumelhart & McClelland 1986: 32, my emphasis).

In comparison to the historical debate between empiricists and rationalists, which revolved around the status of *a priori* beliefs, the debate between cognitive scientists who professed empiricist and rationalist inclinations was centered around the issue what computational models best describe cognitive architecture. The stakes were high—the better and more advanced models would be winners in both engineering and cognitive context. Let me explicitly formulate this phase of rationalist vs. empiricist clash:

> (**E**$_{CS}$) The cognitive processes are best simulated and examined by computational models implementing ANNs. These models are designed to minimize the manual specification of rules and to rely on learning algorithms to prompt ANNs to learn from experience, i.e., data.
>
> (**R**$_{CS}$) The cognitive processes are best simulated and examined by computational models implementing symbolic representations and rules for manipulation over representations. These models are designed to produce human-matching task performance based on the list of instructions that encode expert knowledge.

Both (**E**$_{CS}$) and (**R**$_{CS}$) are essentially empirical claim, i.e., the better modeling approach is not chosen through the conceptual analysis of various commitments and arguments offered in favor of either position but through concrete implementations or empirical studies. The 21st century witnesses the connectionist victory at least in terms of engineering success that has been and continues to be commercially exploited. Nonetheless, the rationalist vs. empiricist rivalry between connectionist and symbolic strands of cognitive science as sketched in (**E**$_{CS}$) and (**R**$_{CS}$) has been around for almost four decades, with little to no progress over the status of connectionism in terms of its autonomy and explanatory prospects. Thus, the trajectory of early ANNs was shaped by institutional and sociological reasons for the rivalry with GOFAI, which are not stressed in the literature. These reasons also play a significant role in the contemporary ANN research—post-connectionist models are walking a tightrope between corporate and commercial success and responsible scientific development. Also, while the early shallow ANNs were designed to pep up cognitive science by unpacking the black box of behaviorism, i.e., workings of the human mind, in a biologically plausible way, thus making use of what we know about the brain to fill in the mechanisms underlying cognitive processes, contemporary ANNs have become black boxes themselves, which is used to stress their principled inability to be explanatory about *anything*.

3. POST-CONNECTIONIST MODELS AND DEEP LEARNING: A SOLUTION TO THE PERENNIAL EMPIRICISM VS. RATIONALISM DEBATE?

> *We all know about the arguments that purport to show that our research can never succeed; indeed, nearly every book written by a philosopher begins with an argument that the competing approaches are hopeless. Yet, for some reason, we persist. Somehow, we are only convinced by the philosophical arguments that everyone else's approaches are hopeless.*
>
> —Anthony Chemero (2011: 4)

3.1. Rationalism and Empiricism of the 21st Century: Post-Connectionist Models on the Battlefront

The breakthrough in the connectionist paradigm came in the last two decades with the availability of the gargantuan training data sets (e.g., *ImageNet* and *WordNet*, see Deng et al. 2009) and greater computational power, which made it possible to advance large-scale big data training of complex, multilayered ANNs. This large-scale training of multilayered ANNs is called *deep learning* — a startling engineering twist that allows for the transformation of raw data into vector-space representations from which the classifier detects a pattern at a higher, more abstract level (LeCun, Bengio & Hinton 2015, *cf*. Hinton, Osindero & Teh 2006).[32] The higher the level, the more likely it is that deep ANN (DNN) will identify the most relevant aspects of input for the cognitive task at hand. Terry Sejnowski, one of the leading figures in DL (along with the previously cited *il trio fantastico* Yann LeCun, Yoshua Bengio and Geoffrey Hinton), describes this as "learn[ing] from data the way that babies learn from the world around them, starting with fresh eyes and gradually acquiring the skills needed to navigate novel environments" (2018: 3). The new generation of post-connectionist models can match or even outperform human experts in many tasks, including abstract strategy games like Go (Silver et al. 2017) or medical diagnosis (Zhou et al. 2021) thanks to processing large datasets, more often in an *unsupervised* rather than supervised manner. New *benchmarks* are being envisaged to deepen and precisely evaluate performance comparison between DL models and humans.[33] Furthermore, these models have vast industrial and commercial usage. All major tech

[32] DL is, essentially, a branch of ML. ML is, essentially, a branch of AI. AI is constitutive of cognitive science along with other parts of hexagon, namely philosophy, psychology, linguistics, neuroscience and anthropology. By extension, thus, DL and ML are used as a tool in cognitive science for developing and testing hypotheses. In what follows, I explore whether DL *is* and whether *it should be* developed in both theoretical and technological sense via the existing results and/or frameworks in cognitive science (or the other parts of the hexagon, most notably linguistics).

[33] Benchmarking in ML amounts to designing evaluations of algorithms to validate a new approach to modelling practice through datasets that take some of the following forms: (i) real-world data, (ii) synthetic data (especially in cases when privacy considerations are at the forefront), or (iii) artificially generated toy data (Torfi et al. 2021: 7).

companies such as *Google*, *Microsoft*, or *Meta* (former *Facebook*), have invested considerable funds in recruiting academics and think-tanks to boost research and development of ANNs as well as benchmarks through which the capabilities of ANNs are being tested.[34]

Ironically, however, their *differentia specifica* is no longer a biological flavor — engineers seldom think whether some architectural feature or learning algorithm is biologically plausible and work under the motto "If it works, don't mess with it." As I elaborated in Sect. **2.3.**, connectionism has been inspired and facilitated by the polyamorous relationship between neuroscience, cognitive science, and AI. For this reason, many researchers who remember the taste of a revolutionary bouquet from 1986, point out that neuroscience and cognitive science can play a vital role in building (more) advanced post-connectionist models, especially those that could aspire to general intelligence (Kiela et al. 2016, Hassabis et al. 2017, Ullman 2019). In other words, pleas for returning to biological plausibility (and body in general) are louder than ever, albeit stem from cognitive and neuroscientists. On the other hand, for engineers, the cash revenue from various implementations of post-connectionist models is the new relevant flavor.[35]

The most pertinent issue in assessing the biological plausibility of post-connectionist models is the type of learning algorithm that modelers choose along with the type of ANN architecture, parameters, hyperparameters (viz., parameters governing the learning process), and the quality and quantity of dataset. There is scarce evidence that error backpropagation algorithm has anything to do with how synaptic connections between biological neurons actually process signals despite some attempts to show that at least an approximation of such algorithm can be detected (see Lillicrap et al. 2014). First, backpropagation is computed linearly, as opposed to biological neurons which make use of both linear and non-linear computation. Second, biological neurons' communication is described by the stochastic binary values of action potentials or neuronal spikes, whereas backpropagation rests on single, static, continuous values.

Deep learning engineers have mostly agreed that more biologically plausible alternatives are needed and proposed to either develop novel ANNs that would mimic neuronal spiking (Bengio et al. 2016, Tavanaei et al. 2019), or novel learning methods that would be akin to ways how animals grasp the world around them, such as reinforcement learning inspired by the Pavlovian conditioning model (Zambaldi et al. 2018). Recently, deep reinforcement learning has been an excellent example of the successful remarriage between neuroscience and AI. The whole point of such learning method is to train ANN to interact with the environment through a planning algorithm and given the observation that it will receive a reward upon producing output (or punishment — should it fail to produce

[34] For instance, *Siri Alexa*, and *Cortana*, intelligent virtual assistants, represent concrete implementations of DL. When it comes to LLMs based on DL, which have been implemented in chatbots, *Microsoft* has partially funded *ChatGPT* and fully developed *Sydney*, *Google* has funded *Bard*, and *Meta* is currently working on its *Galactica*.

[35] Numbers don't lie: *OpenAI*, a *Microsoft*-supported company that launched *ChatGPT* (recall Introduction), is expected to generate around $200 million in revenue by the end of this year and as much as $1 billion by the end of 2024 (Dastin, Hu & Dave 2022). Of course, some scientific applications of DL need not be linked to straightforward financial gains but can equally be indifferent towards biological plausibility of architecture or learning algorithm, such as, say, protein folding models AlphaFold2 and RoseTTaFold that predict functional and accurate structure of a protein molecule from its linear amino-acid sequence (for an overview see Eisenstein 2021).

the adequate output). A DNN is then used to train the so-called Q-network to predict the total reward that can be expected to receive after taking a particular action by relying on the Deep-Q-Network algorithm (François-Lavet et al. 2018).

In the rest of this Sect., I review the state-of-the-art ANN architectures[36] and pave the way for comparing connectionist and post-connectionist models in terms of their theoretical commitments. A word to the wise —we *do* need a full spectrum of different ANNs trained through DL to account for the linguistic competence *in toto*. I also tackle the elephant in the room, i.e., the black box problem that undermines the role of DL-based models and cuts across the traditional notions of explanatory and predictive power in the philosophy of science. As for the architectonic of the whole Ch., the main aim of this Sect. is to offer a general overview of the methodology surrounding post-connectionist models so that I can proceed to assessing their status within the heated Empiricism vs. Rationalism debate that is transcending the disciplinary divide since linguists, cognitive scientists, neuroscientists, AI engineers, and philosophers all have *something* to say about DL. From Sect. to Sect. of this Ch., I am gradually *zooming in* on DL models for NLP and LLMs which represent the main point of contention between Empiricists and Rationalists of the 21st century.

Computational Architectures for the 21st Century

Deep Convolutional Neural Networks

Demis Hassabis, the head of *DeepMind* (a Google-owned AI research laboratory), rightly points out that

> "[r]eading the contemporary AI literature, one gains the impression that the earlier engagement with neuroscience has diminished, [h]owever, if one scratches the surface, one can uncover many cases in which recent developments have been guided and inspired by neuroscientific considerations" (Hassabis et al. 2017: 247).

Hassabis had in mind the deep convolutional networks (DCNNs or CNNs) while making this remark. These ANNs represent a typical example of three familiar trends in connectionist paradigm: (*a*) their architectural features were directly inspired by neuroscience, specifically research on mammalian visual cortex, (*b*) being implemented in models for computer vision, their distinctive success is in line with the historical success of shallow neural networks for lower cognitive processes such as perception (recall **2.3.** and **2.4.**), and (*c*) the initial academic interest in DCNNs has quickly transformed into a lucrative commercial venture thereby taking precedence over further scientific development and their usage. DCNNs were the main vehicle of the deep learning renaissance in the past two decades, albeit their role in scientific research is rarely noticed even though it seems that this architecture has paid back its intellectual debt to neuroscience and cognitive science.

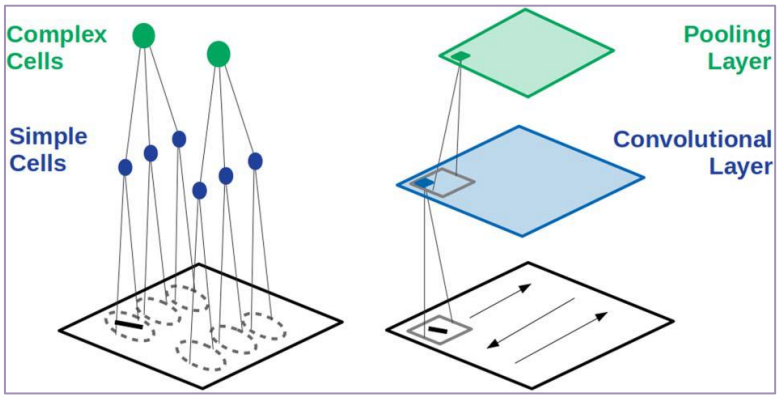
David Hubel and Torsten Wiesel (1959) conducted a series of experiments on cats and kittens as models of the human (or mammalian) visual system by recording signals of

[36] It goes without saying that I do not intend to offer an exhaustive list of all ANN architectures that are currently being used in AI but to focus only on those that figure prominently in philosophical arguments and/or relate closely to simulating linguistic capacities, which is of interest for my thesis. Similar choice of ANN architectures, albeit presented in a less detailed way, can be seen in Torfi et al.'s (2021) survey of NLP in the context of DL.

single neurons and thus creating a detailed map of visual cortex. They discovered that neurons in the early stages of primary visual cortex respond strongly to simple patterns (e.g., bars of particular orientation) but dismiss other more complex patterns. The neurons in later stages are "enlisted" to deal with complex patterns and ignore simple ones. Thus, they distinguished simple from complex cells in the primary visual cortex— simple cells generally have local receptive fields and react to oriented edges, while complex cells, presented also in the secondary visual cortex are organized in a hierarchical manner and remain invariant despite distorted input signals (Hubel & Wiesel 1962, 1963).

Fukushima (1980) constructed *Neocognitron*, a multi-layered ANN inspired by Hubel and Wiesel's findings regarding simple and complex cells. This connectionist model served for handwriting recognition and implemented a first prototype of CNNs, or better yet, both early CNNs and *Neocognitron* share the similar architectural features that make them successful in visual input classification (Rawat & Wang 2017: 2358). Let me now spell those architectural features. First, CNNs are feedforward ANNs (their signal processing is unidirectional) with a biological flavor. Second, CNNs are usually trained through backpropagation — LeCun (1989) was first to apply such trained CNNs to real image classification problems, viz., the classification of zip codes. Third, both early CNNs and *Neocognitron* have simple cells and complex cells.

The early CNNs were sought-after because they relied on a small number of parameters and relied on spatial topology of the data (Rawat & Wang 2017: 2359). However, in time, the datasets and the number of parameters became larger, CNNs deeper, and architectural features more flavorsome, which allowed for avoiding the issue of overfitting. At first, engineers were reluctant to go with the solution of adding more layers to CNNs since this methodological choice is computationally expensive, whereas shallow ANNs are cheaper and easier to train albeit not as accurate as deep ANNs are (Rawat & Wang 2017: 2372). S-cells and C-cells are akin to convolutional and pooling layer in DCNNs (**Img. 1**). These layers are comprised of units computing different activation functions as opposed to shallow CNNs in which all units computed sigmoidal function. Convolutional units, as their name suggests, are activated through *convolution*, i.e., a linear algebra operation that modifies perceptual input (e.g., pixels) in such a manner that some values are favored over others (Buckner 2019: 4). This essentially means that convolutional layers are tasked with detecting relevant features.



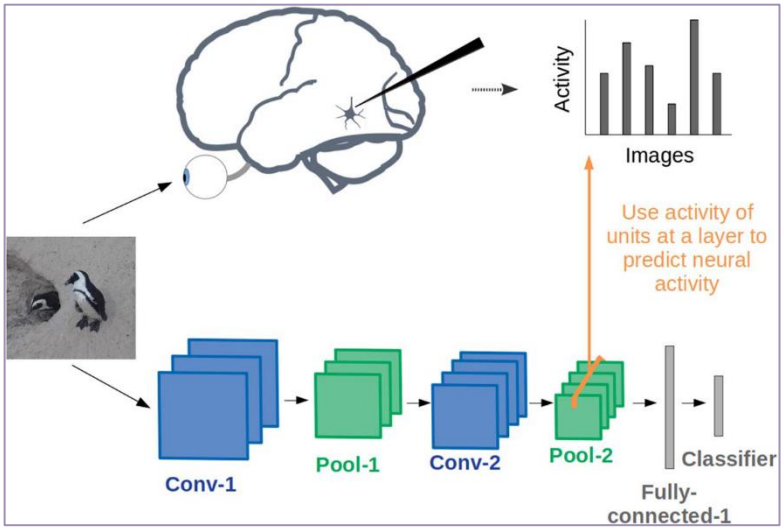
Img. 1 Originally from Lindsay (2021: 2028). The correspondence between Huebel & Wiesel's division of cells and architectural features of CNNs

Specific units called *kernels* amplify shadings or contrasts present in perceptual input and pass on the information to *rectified linear units* which are activated only when the result of the convolution reaches a threshold, i.e., when it is indicated that the relevant feature is found at a particular vector space location (Buckner 2019: 4). The next stage of processing begins with the kernel signal and output from rectified linear units being passed to pooling layer. In this layer, a feature is being detected in all its distinct locations and positions and a down-sampled feature map or representation is being created. The function of the pooling layer is to make such representation invariant to local translations of perceptual input. This is usually done thanks to the max pooling function which calculates the maximum value for each part of the feature map.[37] This function basically forces the network to decide what feature is most salient and most likely to be found at a particular location in the feature map (Buckner 2019: 7). Important thing to note here is that these units are sparsely connected as opposed to fully connected units in any shallow NN of the *Golden Age*. This means that deeper layers take input from nearby units that have overlapping spatial and temporal receptivity from the previous layers (Buckner 2019: 7). The upshot is to obtain through three functions, namely convolution, rectification, and max pooling, a transformed representation of perceptual object or image in the input. The processing flow—from n convolutional layer to m pooling layer, from n_{+1} convolutional layer to m_{+1} pooling layer—ends after many sequences, when, finally, the information is being directed to fully connected classification layer, where labeling of the object or image happens. This last phase is particularly vulnerable to overfitting: DCNNs may learn to simply memorize mappings between objects or images and labels in an exceptionally large training dataset. This forced AI engineers to envisage explicit and implicit regularization techniques. For instance, they add some noise or shifting images to make the DCNNs robust enough to handle such perturbations (Buckner 2019: 8).

Around eight benchmarks mushroomed from 2006 to 2015 to evaluate DCNNs on image classification tasks — and DCNNs were acing all of them (Rawat & Wang 2017: 2368-2369). In 2012, an eight-layered DCNN "AlexNet" (named after its creator Alex Krizhevsky whose Ph.D. supervisor was none other than Geoffrey Hinton) won the *ImageNet* challenge (see Krizhevsky, Sutskever, & Hinton 2012). The challenge required that an ANN be able to classify online images efficiently and relatively accurately from such a large dataset into myriads of object categories.[38] This success caught the attention of neuroscientists who conjectured that the basic features of the visual system could be simulated and further explored through DCNNs. Hence, even though AI engineers were not really aiming for increasing biological plausibility or even paying any particular attention to the neuroscientific legacy of *Neocognitron* and shallow CNNs, neuroscientists have found DCNNs to be their Excalibur — they finally had the chance to validate and further analyze results and data stemming from systems and computational neuroscience.

[37] Max pooling function is not the only available function to the modeler. Rather, one can choose between, say average pooling and max pooling. Average pooling calculates the average value of each part of the feature map so that all parts are equally processed. It is up to a modeler to decide whether the point is to identify the whole object or image or only the most relevant features.

[38] ImageNet challenge was established by an expert in computer vision, Fei-Fei Li, now a computer scientist at Stanford, and held regularly from 2010 to 2017. In seven years, the winning accuracy in classifying images augmented from 71.8% to 97.3%, thereby surpassing human abilities and further promoting the idea that the success of ANNs is proportional to the availability of larger datasets (Gershgorn 2017).



Img. 2 Originally from Lindsay (2021: 2028). Comparison between processing in human brain and CNNs

Grace Lindsay (2021) has called to attention the usefulness of DL in general and DCNNs in particular for neuroscience. In Img. 2 one can witness how the activity of different layers in DCNNs can predict the activity of biological neurons. As Lindsay (2021) rightly remarks, at first it seemed that DCNNs were excellent for explaining how later visual areas (V4 and inferior temporal gyrus) contribute to object recognition, especially last and the penultimate layer, but in the past few years even early-to-middle layers can predict activity of earlier visual areas (V1). Furthermore, DCNNs can produce optimal stimuli for biological neurons thereby providing neuroscientists with tools to control neural activity of primate brain (Bashivan, Kar & DiCarlo 2019) which strengthens the conjecture that these DNNs share some fundamental architectural features with our visual system that allow for perceptual similarity judgments and object recognition. This is the main reason DCNNs were the first weapon to draw for the vindication of empiricism in cognitive science. Their biological plausibility along with predictive success suggests that the state-of-the-art connectionist architectures *can* legitimize domain general mechanisms along with sensory experience taking precedence over innate rules. As I will be showing in the next Sect., this will be Cameron Buckner's key point in developing and defending moderate empiricism, whereas in Sect. 4.3., I will be further developing that point by showing how multimodal DNNs, constituted partly by DCNNs, can account for semantic competence.

The naysayers could, however, beg to differ since despite the grand claims about biological plausibility, DCNNs are *texture-biased* whereas humans and other primates are shape-biased (Geirhos et al. 2019). This essentially means that these ANNs classify perceptual input by relying on the texture present in the input rather than shape when their performance is compared to human performance on the same perceptual task. Moreover, *adversarial examples* have also cast shadow over the success of DCNNs (as noticed for the first time in Szegedy et al. 2013). Adversarial examples are micro modifications to perceptual input that lead to wrong labeling in DNNs but allegedly have no impact on human labeling since such modifications are imperceptible to humans. Thus, such a modification to, say, an image of a panda may provoke a DCNN to erroneously classify it as an image of a gibbon (Goodfellow, Shlens & Szegedy 2014), or generate the so-called rubbish images (Nguyen, Yosinski & Clune 2015). Buckner (2020) has recently argued that instead of lamenting over DNNs susceptibility to adversarial examples, it would be better

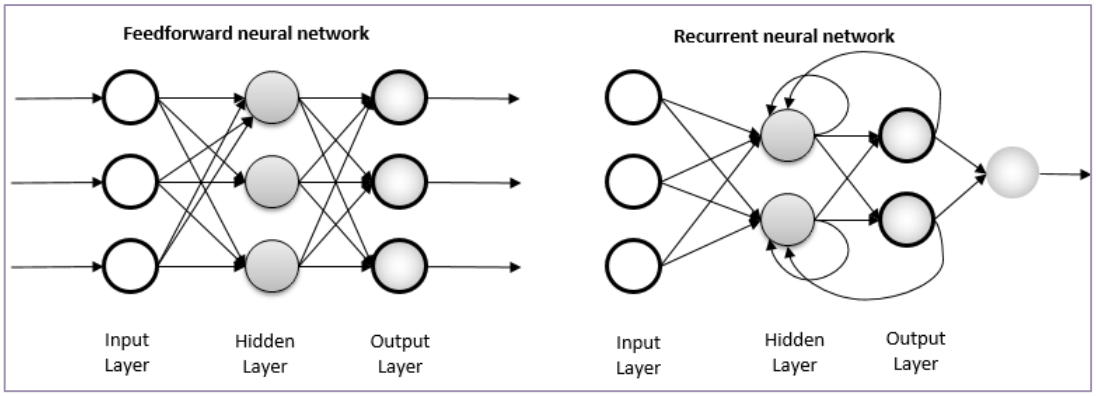
to regard at least some of these examples as artefacts containing predictive information which would not be available otherwise. In a nutshell, artefacts are systematic non-robust patterns that may entail incorrect inferences from data unless we understand their origin and then either cancel their influence on further data processing or use them as legitimate predictions. Perhaps, DCNNs discover intricate interactions that are beyond our perceptual apparatus, which, in turn, allow them to outperform humans, and in that sense, not all adversarial examples are blunders. As Buckner elsewhere wrote:

> "If these categorizations are not necessarily blunders, the ability of [DNNs] to detect the features (...) should no more be counted against their candidacy for intelligence than the ability of Einstein to see things others did not in the equations describing gravity and black holes" (2021: 17).

Long Short-Term Memory Networks

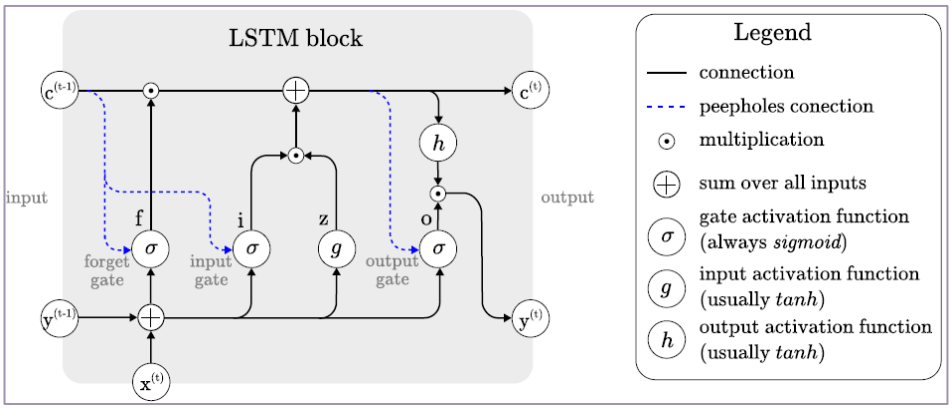
As I have described earlier, ANNs compute numerical functions, and in the case of NLP, input and output units encode words in small and dense vectors that have nonzero values. In this way, similar vectors can be assigned to words frequently appearing in similar contexts, for instance, values (1, 2.3, 3) designate chair whereas values (1.1, 2.3, 2.9) designate armchair, whereas (2.2., -4, 3.1) designate otter. In this way, a manifold of distributed vector representations, or word embeddings, emerge in order to complete a specific NLP task that ANN was assigned to do. The word embeddings are usually learned via gradient descent and backpropagation, although the specific manner of their processing depends on architectural features of ANNs.

Jeffrey Elman introduced SRN in his seminal papers (1990, 1991), in which he reported training a novel type of connectionist model on artificial language sentences. The model was successful in exhibiting emergent lexical classification, i.e., how sentences can be divided into their constituent parts such as nouns, transitive, and intransitive verbs. SRN and any RNN that was further developed for NLP works in the same way: previously hidden vectors in the hidden units' layer are used as additional input when predicting the next word in sequence. Thus, as opposed to a FNN in which there is no cycle of information flow, there is a cycle of information in the form of recurrence in an SRN or RNN (**Img. 3**). As Skansi (2018) rightly points out, SRNs were indeed a milestone in AI research since they successfully grappled with a prejudice that natural language processing is a stumbling block for the connectionist paradigm.



Img. 3 Originally from https://www.researchgate.net/figure/315111480. Architectural differences between FNNs and SRNs.

As soon as 1997 (that is, only seven years after Elman's seminal paper where he introduced SRNs), German engineers, Sepp Hochreiter and Jürgen Schmidhuber, reported designing and training a novel type of networks based on recurrent architecture and named them LSTMs.[39] Their paper is now considered one of the most-cited papers in AI research. Essentially, LSTMs are SRNs on steroids because they are designed to remember information for longer time, i.e., to handle long time lags that SRNs could not. LSTM are composed of many recurrent subnetworks that serve as memory blocks (**Img. 3**). While SRNs operate through a single connection from one unit to another, LSTMs are endowed with memory cell state (C_t) and filters, or gates, that constitute each memory block. The role of the gates is to determine whether information should be removed or kept in the C_t in order to perform a task. Thus, if it is vital for the task performance that NN removes the information, then forget gate is called to the rescue.[40] This gate should "decide" how much of weighted input and previous hidden states should be in the network's memory. The input gate is "entrusted" with adding information to C_t, and "decides" how much of weighted input should be saved. Finally, the output gate serves as a function mechanism which "decides" about the crucial parts of C_t. The gates, along with continuous flow of data processing, allow LSTMs to avoid vanishing/exploding gradient.



Img. 4 Originally from Van Houdt, Mosquera & Napoles (2020: 5932). Typical LSTM architecture. Reproduced with permission from Springer Nature.

The main advantage of LSTMs over SRNs is, in fact, their successful dealing with vanishing/exploding gradient. Moreover, the issue with vanishing/exploding gradient was the main impetus behind developing LSTMs. In fact, Sepp Hochreiter started analyzing vanishing gradient problem for his BSc thesis supervised by Schmidhuber back in 1991, six years before their seminal and highly cited paper on LSTMs.[41] Thus, the incentive for technological innovation came from the refusal of an AI researcher to come to terms with current methodological constraints. In the rest of this Sect. and Ch., I argue that this quite frequent moment in the brief history of connectionism is one of the key reasons why we should regard connectionism as a future-oriented research program. Recall now that due to the amount of the network's (hidden) layers and propagation of error through the time

[39] In 1995, Hochreiter and Schmidhuber published a technical report in which LSTMs appear for the first tim, whereas the peer review process prolonged the publishing of a scientific paper devoted to LSTMs to 1997.

[40] Forget gate was added some two years after Hochreiter & Schmidhuber (1997) in Gers, Schmidhuber & Cummins (1999). Later, a variation on forget gate appeared in Cho et al. (2014) as *gated recurrent unit*, now widely known as GRU.

[41] In a blogpost, Schmidhuber symbolically called this year *Annus Mirabilis* at TU Munich.

loop, activation functions governing weight updates squish large input into smaller vector spaces between 0 and 1. This results in the inability to propagate relevant information and, consequently, the unstable behavior of a network that is being trained: the gradient either exponentially diminishes to 0 or increases above 1 and explodes. LSTMs are up to the challenge here. The gates handle the amount of lost gradient since their activation values differ at *t, t_1, tn*, and these values are *learned* functions based on the current input and myriads of hidden layers.

LSTMs can be trained on large sets of textual data through DL algorithms and are mostly used for sequential tasks such as machine translation, speech recognition, robot control, musical and language processing, etc. However, LSTMs are often used alongside CNNs to optimize task performance resulting in post-connectionist models with multimodal ANNs (for a brief overview see Van Houdt et al. 2020: 5948). As I will argue in **4.3.**, multimodal models are crucial for simulating semantic processing, and, thus, providing us with a patchy and messy account of linguistic competence, which better reflects the nature of the language faculty than idealized and normatively "zipped" account that rationalists promote. When it comes to their industrial application, all three giants among tech companies, namely *Apple*, *Facebook* and *Google* use LSTMs for their intelligent virtual assistants and automatic translation of messages within applications.

It is also worth noting that SRNs have evolved into multilayered RNNs, which are still used in computational linguistics and psycholinguistics alongside or in comparison to LSTMs. These RNNs can have strong structural priors, such as gates, encoders, decoders, and attention mechanism. Unlike SRNs that were based on data stemming from artificial languages, RNNs, just like LSTMs, process natural language corpora in a sequential manner and are used for describing grammar learning. The difference between these two similar architectures is the length of statistical regularities they can capture. LSTMs, due to their architectural features, excel at capturing long distance statistical regularities. Like LSTMs, RNNs are also used within multimodal post-connectionist models as decoders to produce linguistic output, along with DCNNs used as encoders producing visual input. For instance, models for image caption generation learn to describe the content of images by taking images (usually from ImageNet dataset) as input and produce natural language paragraph or sentence (Xu et al. 2015, Krause et al. 2017). Other domains of application of such models are visual question asking and answering (Wang & Lake 2021), instruction following (Ruis et al. 2021), and labeling video frames (Yeung et al. 2018).

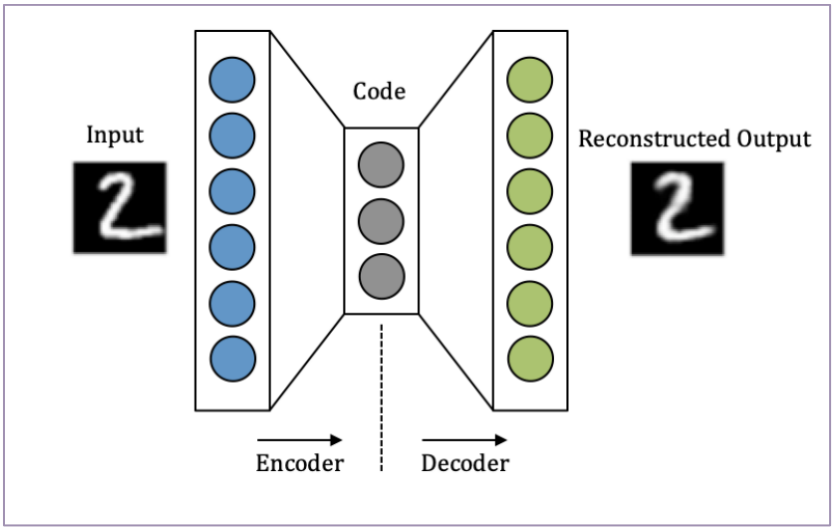
Autoencoders and Transformers

Both autoencoders and transformers are state-of-the-art architectures in the post-connectionist paradigm. Fuss aside, both can be considered yesterday's news among the real AI connoisseurs. Ballard (1987) proposed something remarkably similar to current autoencoder architectures, as well as one of the pioneers of DL, Yann Le Cun, in his doctoral dissertation defended the same year at the University Paris 6.

As opposed to previous NNs that mostly presumed training through supervised learning, autoencoders crunch data mercilessly through unsupervised learning. The difference between supervised and unsupervised learning was introduced in Sect, **2.3.**, but I will rehearse it here given that this distinction will figure prominently in the next Sect.

Supervised learning is a way to train ANNs through labeled data by mapping input in the form of a vector to output value in the form of pre-defined label. The mapping function is inferred by an algorithm that a modeler has previously chosen. Testbed for supervised learning is the generalization to novel examples, i.e., the prediction of labels to previously unseen data. This is nowhere trivial thing to do since previously unseen values come with arbitrary output values. Hence, the algorithm always comes with at least some inductive biases, that is, assumptions that "nudge" ANNs into favoring correct predictions over incorrect ones. Unsupervised learning, on the contrary, amounts to making sense of data without using labels or specifying parameters and hyperparameters. Briefly, a modeler does not have to intervene, but rather to analyze patterns that ANNs have produced through unsupervised learning.

Autoencoders are much the same as FNNs since they are also three-layered, but their task is to efficiently recreate the input in an unsupervised manner, by ignoring noise in unlabeled data (Skansi 2018). This, in turn, means that both input and output layer must have the same number of units, whereas in the hidden layer, also called encoder, there are fewer units than in the previously mentioned layers (Img. 5). All subtypes of autoencoders serve for efficient preprocessing of data regardless of their architectural differences.[42] The modelers are interested in the activation values of the hidden layer because these values will be used as input in a bigger ANN. The point is to compress data through an autoencoder, so that data can be uncompressed in the most suitable manner to match the input of the new ANN. Autoencoders are *generative* models, which means that they can be used for creating training data for bigger ANNs, so multiple autoencoders are usually stacked within a deep ANN. This means that autoencoders can equally be a part of task performance in the domain of computer vision (i.e., stacked within a DCNN), or in the domain of NLP (i.e., coupled with a transformer).

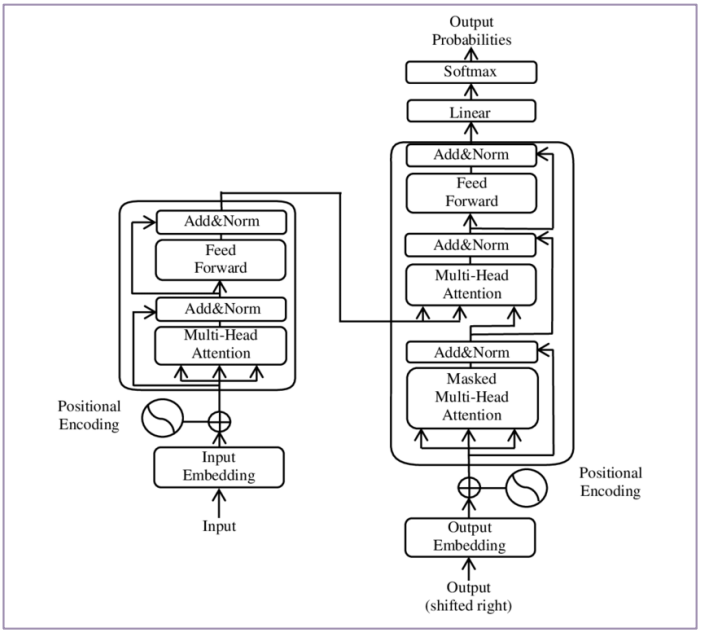


Img. 5 Originally from Torfi et al. (2021: 3). Typical autoencoder.

[42] The most frequent classification of autoencoders is into simple, sparse, denoising, and contractive (Skansi 2018: Sect. 8.2). I have described simple autoencoders. Sparse autoencoders have restricted number of units in the hidden layer (e.g., double the number of units in the input layer), denoising autoencoders handle noisy input as well, whereas contractive autoencoders have additional explicit regularization techniques for handling noisy input.

Transformers are avant-garde ANNs for NLP since they are implemented in LLMs like BERT (Devlin et al. 2018), GPT-2/GPT-3 (Brown et al. 2020), or multimodal languages like DALL-E (Ramesh et al. 2022), as well as the main drivers of success of chatbots like ChatGPT. As in the case of auto-encoders, even though they were thrust in the limelight six years ago, transformers sprang up by the end of the 20th century. In 1992, Jürgen Schmidhuber, the brain behind LSTMs, had already published a paper on fast weight programmers, which were akin to contemporary transformers with attention heads. Nonetheless, Vaswani and colleagues (2017), a group of AI researchers working at *Google Brain*, re-introduced transformers to develop pre-trained language models that should be fine-tuned for specific (and ultra-commercial) tasks.

The leap from learning mere patterns from a large body of text to obtaining a general capacity for processing language and, thus, learning syntactic and semantic structure (to at least some degree), represented a revolution in NLP. The revolution was made possible because transformers incorporated a secret ingredient besides having computationally powerful (auto-)encoder/decoder architecture, namely attention heads arranged in layers, in similar manner as regular units (**Img. 6**). However, unlike regular units, attention heads perform distinct operations—they allow model to remember multiple words of input, which, in turn, amounts to in-context learning. In this way, transformers "focus" on specific parts of the input while processing large amounts of corpora *at once*. Recall that RNNs and LSTMs are trained to predict the next word in the sequence given the other words that were processed. Transformers can "remember" all tokens of the word because they have processed every token position (i.e., context) in the training dataset, thereby easily accounting for polysemy.



Img. 6 Originally from Jia (2019: 012186/4). Typical transformer architecture.

More precisely, encoder consists of an FNN and attention mechanism that receives information about isolated word embeddings, i.e., low-dimensional vectors that build matrices of word co-occurrences in the training dataset. The next step is positional encoding: a fixed-size vector captures relative position of isolated embeddings which helps preserve their identity while signal processing spreads through the rest of the transformer

layers (Lake & Murphy 2021: 7). Decoder, much like encoder, consists of an FNN and two rather than one attention mechanism so that information about position can be used as input for decoder. The more stacked with encoder/decoder architecture transformer is, the deeper context can be processed. Through each transformer layer, word embeddings are updated with novel context, which results in generating a whole text or paragraphs based on a single prompt. Transformers have outperformed RNNs when it comes to self-paced reading, which made Merkx & Frank (2021) doubt the assumption that recurrent mechanism lies behind sentence processing. In other words, they are—arguably—more psychologically plausible than RNNs. Of course, such claims should be further evaluated in empirical and engineering terms.

Transformers are notorious for having billions of parameters and hyper-parameters fine-tuned after unsupervised pre-training that helps setting up the initial set of parameters. RNNs and LSTMs are usually trained in a supervised manner that requires manual labelling of data, which makes the mission of creating LLMs out of these ANNs time-consuming and costly. However, even though transformers have an advantage over RNNs and LSTMs regarding the financial and computational feasibility of developing and training LLMs, they are also less transparent with respect to their inner machinery given the exorbitant number of parameters. Thus, GPT-2 has as much as 1.5 billion of parameters and is trained on 45 million Reddit threads and pages (Radford et al. 2019), while BERT has 304 million of parameters and is trained on *cca* 3.3 billion Wikipedia pages and a large set of unpublished books (Devlin et al. 2018). GPT-3 has 175 billion parameters and the largest out of the four training sets that amounts to 410 billion tokens from the internet. GPT-4 has been released in March 2023 and is presumably even larger, but technical information about its size and training sets have not been disclosed yet.

Why go large, though? It seems that the larger any LLM is, the further is from the way humans learn. We are, allegedly, capable of zero-shot and few-shot learning—we do not need the gargantuan amount of data to successfully generalize to previously unseen exemplars.[43] Furthermore, LLMs are trained via algorithms belonging to a class of deep reinforcement learning, which can make us doubt that they genuinely learn from experience instead of just chasing the reward signal (say, like some sort of cyber-Pavlov's dog, or better yet cyber central nervous system of Pavlov's dog). Buckner (2021) considers these criticisms as being instances of *anthropofabulation*, i.e., comparative bias which makes us score nonhuman performance against inflated view of our competence due to being empirically uninformed about our own cognitive processes. By closely examining studies in developmental and social psychology, Buckner points out that we can be equally charged with the same transgressions as any other data-hungry, reward-chasing DL model.

[43] In ML, zero-shot learning refers to cases when a model is trained to recognize and classify data that it has never seen or encountered during the training phase, while few-shot learning refers to the cases of training with a limited amount of data. The key point behind both types of learning is to transfer knowledge from the known classes to the unknown ones because in that way the model exhibits generalization skills and flexibility in performance. This part of ML was directly inspired by zero-shot in humans (see Palatucci et al. 2009), given that a body of empirical evidence suggest that humans have the capacity to easily generalize knowledge to unseen and unknown objects or concepts thanks to either their algebraic mind (Marcus 2001) or semantic associations (Rogers & McClelland 2006). The interpretation of evidence is, obviously, dependent on what type of cognitive architecture one endorses—symbolic or connectionist.

Thus, we are faced with so many training exemplars that common sense usually neglects due to being used to them. For instance, different vantages of the same object are, in fact, distinct exemplars for our cortex, and the role of offline memory is to replay such exemplars or even simulate new ones based on the past experience (Buckner 2021: 13). Without taking this into account, we might wrongly presuppose that an infant exposed to a single object such as a toy engages in zero-shot learning. However, infant grasps and processes all the vantages, features, and other aspects of a single object during the exposure. Furthermore, we are reward-hackers *par excellence*: our "striving and starving" ancestors (helped by the mechanism of natural selection) applied a trial-and-error approach to problem solving in their environment (Buckner 2021: 22). As Buckner argues, this happened through positive reinforcement of some bodily movements rather than others, nervous system that updates its status via multimodal sensory input, sensory organs that are fine-tuned to stimuli of specified range, and brain processes customized by neurotransmitters and hormones. In this sense, tinkering DL models to be less data-hungry or reward-chasing would—ironically—make them less biologically plausible and adequate for comparison to humans. In Buckner's words:

> "Anthropofabulation suggests that humans have some uncanny innate ability to flexibly pursue intrinsically valuable goals in highly diverse environments; but a fair appraisal of modern life would suggest that humanity is not currently doing so well at this particular balancing act." (2021: 23)

My aim in Ch. **4** will be to present computationally feasible domain of transformer based LLM's syntactic and semantic competence, while, at the same time, avoiding the pitfall of anthropofabulation.

As I have sketched so far, contemporary post-connectionist models resemble a patchwork of modeler's decisions, and preferences with respect to types of ANNs, learning algorithms, parameters, and training materials. Also, they seem to serve industrial purposes rather than purely scientific: the progress after the 2000s was fueled by engineering drive rather than the ambition to establish connectionism as the autonomous paradigm for understanding the human mind. Representative industrial and scientific applications of the above described ANNs can be consulted in **Tab. 2**.

Tab. 2 An overview of different ANNs with respect to their task and type of application

Type of ANN	*Representative task*	*Representative industrial application*	*Representative scientific application*
Feedforward (multilayered)	Classification, time series prediction	Detection and prediction of diverse sonar signals	Protein secondary structure prediction
Recurrent	NLP, speech recognition	Text-to-speech systems	Grammar learning
Long Short-Term Memory	Speech recognition, machine translation	Google Translate	Sequence to sequence (seq2seq) learning
Convolutional	Computer vision	Facial recognition, social media tagging	Medical imaging, radiomics (R-CNN)
Transformers	NLP	ChatGPT, Sydney, Bard, Galactica	GPT-2/3/4, BERT, LaMDA, DALL-E 2

It is worth noting, however, that post-connectionist models come with theoretical commitments that are similar to those of the previous era (e.g., representationalism and

empiricism) despite the methodological differences (consult **Tab. 3** for a digest). Cameron Buckner pointed out in his 2019 paper that philosophers largely ignored DL since novel ANNs are just "more of the same— perhaps an important engineering advance, but incremental rather than game changing" (p. 2), thereby almost completely missing the kairos (καιρός). The winds have changed in the past four years, and philosophers are once again engaging with (post-)connectionist paradigm to tackle the perennial issues in philosophies of mind, science, or language through the DL lens (obviously, given the topic of this dissertation).

Tab. 3 Differences between connectionist and post-connectionist paradigm

Connectionist Paradigm	*Post-Connectionist Paradigm*
Shallow NNs (up to 4 layers)	Deep NNs (from 5 to more than 250 layers)
Fully connected	Sparsely connected
All processing units employ sigmoidal activation function	Diverse kinds of processing units employ different activation functions
Scarce number of different architectures	Considerable number of different architectures
Scarce number of parameters and hyperparameters	Considerable number of parameters and hyperparameters
Backpropagation algorithm	Deep learning; Reinforcement learning

So far, different computational architectures have been presented in the light of their task performance which may or may not be specifically linked to human cognitive processing. As it was the case with the shallow connectionism in the 1980s, the usage of DL models in cognitive science comes with the expectation that computational architectures they implement should be illuminative regarding cognitive architecture— despite their decrease in biological plausibility in comparison to the commandments of the PDP Bible. In Milojević & Subotić (2020), we have argued that once post-connectionist models are understood as having an *exploratory* role, *prima facie* concerns, such as their lack of biological plausibility as opposed to shallow connectionist models, can be successfully appeased. Moreover, attributing the exploratory role to these models helps us unite theoretical commitments and actual engineering practices which made possible the rapid progress of DL models in the last two decades. Two things impede such unification. First, engineers sacrifice representational accuracy for boosting predictive accuracy, which is something that philosophers of science frown upon. Any model, it seems, must be both representationally accurate, i.e., faithful enough to reality, *and* have predictive power to control such a reality. Otherwise, models are not empirically adequate, i.e., they are not describing reality and, consequently, we are unable to understand it via such models. Second, due to the proliferation of models implementing all sorts of architectures, and their usage in diverse scientific disciplines, the role of DL models must be defined broadly to cover all different instances in a theoretically consistent and neat way. Both impediments can be resolved through understanding DL models as exploratory *and* idealized models that do not shy away from incorporating distortions to boost computational efficacy. Do bear in mind that I will proceed to defend all these claims only in the domain of cognitive science, neuroscience, and by extension to linguistics once we pass to LLMs in Ch. **4** and that this defense essentially develops argumentation already suggested in Milojević & Subotić (2020).

Collin Rice (2018, 2021) argues that physics and biology are loaded with indispensable distortions whose removal would result in the complete disassembling of the model. The reason for such disassembling can be found in the fact that models do not contain isolated distortions but represent a *holistic distortion*, i.e., they misrepresent phenomena, interactions between phenomena and relevant properties of phenomena (Rice 2018: 2796). Were it not for such "pervasive system-wide distortions," sophisticated mathematical techniques could not have been applied, and modal information pertaining to counterfactual relevant and irrelevant properties of the model's target system could not have been extracted either (Rice 2018: 2809-2811). The same applies to DL models as used in cognitive neuroscience. As Catherine Stinson (2018, 2020a) notes, connectionist models are not constructed for the deduction of precise brain activities. Moreover, they are not constructed for deduction at all —symbolic models, being grounded in Newell & Symon's (1961) physical symbol system hypothesis and inspired by Hempel's (1958) account of deductive-nomological explanation, rest upon the assumptions that predictive and representationally accurate cognitive models must establish a deductive inference from antecedent laws or theories to empirical observations and, thus, provide explicit and transparent explanations. On the other hand, Stinson remarks that connectionist models reflect logic of tendencies, i.e., the upshot is to *discover* and *explore* what tasks brain could do given the specific parameters and input data. As she puts it:

> "First we discover, through a combination of mathematical demonstration and empirical observation, that a certain type of mechanism [...] tends to give rise to a certain type of behavior […]. We then make use of that knowledge to make sense of how brain structures give rise to cognitive phenomena" (Stinson 2018: 130).

The exploratory role of models presumes that models are used as starting points for theory building. In other words, modelers are free to (*i*) generate new hypotheses with loose analogies between model and target system, (*ii*) tinker with model's parameters, and inner functioning in general, to come across novel ways of processing input or designing novel architectures, (*iii*) provide proof of principle demonstrations, (*iv*) assess the suitability of target system, or (*v*) refute an impossibility claim regarding the relation between a model and target system (Cichy & Kaiser 2019: 308, Sjölin Wirling & Grüne-Yanoff 2021). These models are often incomplete, sketchy and contain idealizations and distortions. However, as Stinson argues, they are promising tools for the exploration of *general-level mechanisms* underlying cognitive processes. These mechanisms are instantiated by, say, a DL model and brain, which, in turn, makes it possible to draw inferences from the former about the latter. What matters is that the model and target system share abstract mechanistic structure, rather than representational accuracy. A modeler may toy with architectural features, tinker with the different parts of a model, and, ultimately, see what *could possibly* give rise to some of the tasks we perform every day—from detecting objects to parsing sentences of our mother tongue.

Thinking outside the (Black) Box

So far, the story checks out: there are myriads of different and exotic post-connectionist models, some of them may be considered exploratory, and specifically for the domain of cognitive and neuroscience, this role seems particularly promising. However, do such models *explain* mechanisms besides exploring them? Does tinkering and toying with such models result in promoting any scientific purpose? Surely, you would be right to

expect that post-connectionist models must offer us an explanation as well, if they are to be understood as models of *human* cognitive processing, i.e., as informative regarding, say, our ability to produce and comprehend language.

I will argue in this subsection that the misunderstanding of the role of DL in linguistics, cognitive, and neuroscience (*ipso facto* in the debates pertaining to the nature of cognitive mechanisms) arises when one demands the impossible from post-connectionist models.[44] The result is usually quite anti-climactic: labeling models as "black boxes", i.e., uninterpretable and opaque with respect to their inner machinery amounts to denying them explanatory ambitions in general and undermining the advantages of individual models. Interestingly enough, connectionism was pitched in the 1980s as a computational and theoretical tool that would unpack the mind conceived as the ultimate black box in behaviorism. Thus, in a way, cognitive science has been dealing with the "black box problem" from its genesis albeit not in the fatalistic sense as the contemporary literature on methodology of DL prefers. Moreover, labelling anything as a black box is rather indicative of immature science or an undeveloped research program that is yet to grapple with all the puzzles. Nonetheless, before developing further this point against the generalist claims *contra* explanatory prospects of DL, I will discuss the black box problem in more detail. There are two main reasons for this:

(*I)* Some AI researchers without background in either philosophy of science or cognitive science may find the discussion about post-connectionist models a bit unsettling: it appears that we are bickering about far-fetched ideas whereas the interpretation of models' behavior, on which we build our arguments, amounts to a mere leap of faith. (*II*) Explainable AI (XAI) has been gaining momentum, and its core objectives are becoming a touchstone for AI research in ethical and methodological terms. The upshot is to throw light on "black boxes" in such a way as to make them understandable to humans (at minimum to domain experts, at maximum to end-users).[45]

Thus, it seems that to discard *(I)*, one must address *(II)*, at least in passing. This essentially means that philosophical considerations regarding DL cannot be placed in the grand scheme of things without putting at ease actual modelers that use and develop DL. In other words, before we assess current and propose novel argumentative strategies, we ought to convince the DL community that we *qua* philosophers fully understand the limitations and technical intricacies of their models, and that our interpretations of models' behavior are built on a solid foundation instead of wishful thinking. A particularly useful way to deal with (*II*) is, I believe, through the philosophical analysis of the concept of

[44] In what follows, I quarantine linguistics, cognitive, and neuroscience, and apply philosophical considerations pertaining to explanation only to these fields as I did with claims about the exploratory role of DL models. A more nuanced analysis that would unify different usages of DL in science generally (and the feasibility of such endeavor) is beyond the aim and scope of this thesis. See Bianchini, Müller & Pelletier 2020 and *Royal Society and Alan Turing Institute* 2019 for the assessment of the role of DL in the other scientific fields.

[45] XAI approach has been mostly advanced (and funded) by DARPA with the goal of remaking "black boxes" into "glass boxes" so that end-users can learn when to trust AI systems. Here on the Old Continent, the EU General Data Protection Regulation explicitly states that end-users are entitled to "the right to explanation" of algorithmic decision-making based on or affecting their personal data (European Commission 2016: 71). Strictly speaking, this means that end-users are to be given justification for the decision of AI system fed with their personal data. However, in order to give them such a justification, the domain experts should be able to explain the system's behavior.

explanation as I have mentioned above.[46] In this way, I will try to provide an answer to the worries expressed in *(I)*.

Of course, my intention is not to "solve" the black-box problem but rather to provide a lens through which it is possible to bypass at least some of the roadblocks on the path towards the connectionist account of linguistic competence I intend to develop in this dissertation.[47] Thus, I want to show here how DL models can retain explanatory power *despite* the black box problem and what sort of explanation we can reasonably expect to obtain from such models. Although it seems that there is nothing really controversial in attempts such as XAI that should be regarded as a natural step in the process of racking our brains over complex ML models, somehow there is an impression that the very demand for something like XAI makes ML models dubious or inferior. History repeats itself once again — a tacit comparison to symbolic models is to be blamed here for such impressions that shape our expectations of ML models' explanatory power. Expert symbolic systems in the 1980s and 1990s could provide users with verbal descriptions of their problem-solving strategies, whereas early ML model *PROTOS* (Bareiss et al. 1990) that implemented symbolic architecture could even explicitly explain its actions. These systems were considered transparent and interpretable since AI researchers could effortlessly reverse engineer the reasoning from conclusions to assumptions thanks to the rules governing logical inferences (recall Sect. **2.2.**), and thus *explain* the machine behavior. This is the kind of explanation we *cannot* reasonably expect from ML models. Rather, the way ML models operate suggests we need to re-conceptualize the traditional notion of explanation in the philosophy of science.

When we refer to DL models as black boxes, we want to convey in the most general form that their functions are too complicated for us to comprehend, whether we are AI researchers or mere mortals (Rudin 2019: 206). Nonetheless, what may seem odd is that, at the same time, many DL models have remarkable predictive power. Traditionally, in the philosophy of science, one of the distinctive features of the scientific explanation is predictive accuracy (besides representational accuracy): if a model provides us with

---

[46] This is not the only option available to philosophers, though. Beisbert & Räz (2022) enlist philosophers who approach the black box problem from the perspective of providing conceptual analysis of the notion of *understanding* rather than explanation. However, the key question of one of the examiners that I had to address while defending my dissertation proposal was, roughly, why do we, as philosophers, think that we *obtain* explanations from such models (regarding, say, human cognition) when they, as computer scientists, do not have the slightest idea whether such models are explanatory *at all*. My motivation here is thus to try to provide conceptual means to discard such worry by analyzing what kinds of explanation we, as philosophers, have at our disposal to initiate the dialogue with them, computer scientists, who are striving for the explainaibility.

[47] I am also fully aware that this line of defense does not address the severe social problems stemming from the black box problem. For instance, facial recognition models based on ML impacted high-stakes decisions in criminal justice, resulting in incorrectly incarcerating people, denying parole, or even releasing dangerous lawbreakers, which is evocative of the worst days of phrenology (Stinson 2020b). LLMs, such as Open AI's GPT-3, easily reproduce racist, homophobic, and misogynistic hate tropes unless explicitly censored, and, paradoxically, attempts to de-clutter such models have led to boosting the discrimination against minorities (Xu et al. 2021). For this reason, the pleas for using simplified interpretable models rather than trying to explain opaque models have recently become more prominent (e.g., Rudin 2019). For the time being, my focus will be solely on the methodological aspects of the black box problem. In **Conclusion**, however, I will do my best to draw practical implications of the considerations presented in the thesis, which will be tightly intertwined with the social aspect of this problem.

successful prediction, we have a good reason to believe it fits the data and captures the sought-after understanding of the phenomenon. Thus, it seems that predictive power goes hand in hand with models being explanatory as well. Most methodological accounts of ML have stressed that the connection between predictive and explanatory power is cut in such models. Since DL models are only a subtype of ML, the same should apply to them. However, the upshot is to make such models *interpretable* by unpacking the black box. The way we can grasp the predictive accuracy amounts to detecting the *underlying mechanism* by constructing an internal explanatory model through various analytic techniques as it is widespread practice for ML models within the XAI approach (Gilpin et al. 2019, *cf.* Rudin 2019).[48] This is where the opacity enters the scene.

Following Humphreys (2009), we can define a process as epistemically opaque when any agent *X* at time *t* does not manage to grasp all the epistemically relevant elements of the process. Now, recall that DL models have various parameters adjusted through training (e.g., weights, activations, etc.) and hyperparameters (e.g., general type of architecture, number of layers, choice of activation, learning rate, etc.). Both parameters and hyperparameters lack any assigned meaning. This is quite natural when you think about it —the modeler would have to manually assign meaning to every node and set up their weights, which is an irretrievable loss of time and resources and seems virtually impossible for tasks demanding multiple ANNs. So, how do we trace the underlying mechanism beneath the functioning of DL models? Boge (2021) notes that three steps are critical in this regard:

(a) the conceptualization and approximation of input and output functions,
(b) following (*a*), a modeler establishes the target phenomenon,
(c) through the deployment of background theory, a modeler can connect previous stages to the underlying mechanism.

However, a DL model *per se* may easily lack conceptual apparatus for providing us with full transparency of its mechanisms. First, many different approximations can be used for input and output functions, so we can never be sure whether a model provides us with genuine causal inference (or at least statistical correlation) regarding the target phenomenon, which, in turn, jeopardizes the prospects of validation and generalizability of the model.[49] Thus, the present ambiguity in step (*a*) affects directly (*b*) and, *ipso facto* the last step (*c*). Second, both background theory and conceptualization are out of reach when a

[48] According to Lipton (2018), post-hoc interpretability should be distinguished from interpretability *qua* transparency: the former encompasses what is usually dubbed XAI, whereas the latter refers to the quest of understanding why model behaves in a particular way. Rudin (2019) is quite skeptical about the pursuit for post-hoc interpretability and calls for constructing internal simplified models that would aim for interpretability qua transparency, since post-hoc XAI techniques are not always reliable. My argumentation will hinge on both XAI *sensu stricto* (post-hoc techniques) and XAI *sensu lato* (internal model construction), and when I use the term "XAI approach", I understand it as an approach trying to accommodate the two types of interpretability.

[49] Lapuschkin et al. (2019) dub this the "Clever Hans phenomenon", a phenomenon named after a German horse who could allegedly perform arithmetic operations. Psychologist Oskar Pfungst debunked the marvelous capabilities of Clever Hans by discovering that this was a mere artifact: Hans responded to body cues of his trainer instead of making genuine inferences. When applied to ML models, this phenomenon indicates that the output of a model will be irrelevant and misguided when deployed in the real world, since provided classifications were not based on genuine inferences but rather artifactual associations.

model has an exploratory role and exploits an unsupervised learning method, and neither of the steps can be fixed in this case. Boge's point is that not only do we have to face the opacity coming from the complexity of learning algorithms, but it is nowhere near evident what it is about data that allows DL models to be predictively successful. This means we have to deal with two aspects of opacity — it is epistemically opaque *how* the DL model learns as well as *what* was learned by the model (Boge 2021). Thus, in his view, opacity cannot be addressed *in principle*.

Now, although Boge paints a grim picture of DL models when it comes to their interpretability, I still think that his pessimism can be avoided even though the crux of argumentation holds. Let me return to the definition of epistemic opaqueness. It seems that what counts as an opaque, epistemically relevant element of the process will depend on the type of agents who differ with respect to their initial epistemic status. Zednik (2019), following Tomsett et al. (2018), classifies agents into five groups: operators, creators, examiners, data subjects, and decision subjects. Operators specify the model's input and receive output, and creators are, to put it simply, modelers. Examiners are responsible for risk assessment regarding various laws and regulations. In contrast, decision subjects and data subjects are end-users directly affected by the model or whose personal data is part of its training materials. The five groups of agents are faced with the epistemic opacity of the model from different perspectives.

Carlos Zednik makes an excellent move by linking five groups of agents to distinct levels of explanation through a crude analogy with Marr's levels (1982), namely computational, algorithmic, and implementational. Marr introduced the three levels to analyze information processing: the first level is concerned with computational problems that information-processing system ought to solve, the second refers to the exact processes and format of representations needed for problem solving, and the third specifies physical implementation of this system into hardware or biological body. In Zednik's view the first level can be taken to delineate *what-* and *why-questions* that XAI can tackle: what representational vehicles govern a model's behavior, and thanks to which representational content such behavior can be interpreted?[50] Zednik (2019) notices that operators would be more interested in trying to come up with answers to what-questions, while examiners, data, and decision-subjects would be concerned with answers to why-questions. Algorithmic and implementational levels are related to *how-* and *where-questions* on which XAI techniques provide answers: how model connects the input to output, and where exactly this connection is being realized. According to Zednik, creators will be wrestling with answers to these questions by searching for an abstract mathematical description of properties and the physical realizability of such properties.

[50] Zednik interprets the first and second level in similar terms: he takes computational aspect in the first level to be intrinsically linked to representational content, which is then only specified in the second level. This aligns with the general tendency to consider any discussion of computation as being semantic in the spirit of traditional symbolic cognitive science. However, Ritchie (2019) argues that philosophers, like Zednik, often disregard the fact Marr had neuroscientific motives and adhered to neuroscientific methodology, therefore his computational theory which occupies the first level is concerned only with the mathematical specification of the computational problem that the information-processing system should solve. Ritchie's point aside, I still think that Zednik's classification has merit and his reliance on Marr should be understood as a loose heuristic.

XAI encompasses a plethora of post-hoc analytic techniques for rendering models interpretable (e.g., text-generators, visualizations, nuanced statistical analyses), and for some of them, Zednik convincingly shows that they can be useful for answering these diverse questions. In this way, he paves the way for developing a normative framework for XAI that would serve to show that the black box problem is not insoluble — at least for some agents, given their epistemic status and epistemic demands. For instance, input heatmapping, i.e., the process of highlighting the model's features in the input layer pertinent for the output, provides different agents with answers to both what and why questions.[51] Operators' epistemic demands can be appeased since one can trace an error in the output of the model trained for image classification to input features endowed with "higher responsibility" (e.g., wrong sampling of cyan color). Even creators can profit from this technique since it will become apparent where the necessary adjustments must be made. Moreover, in cases when highlighted features bear resemblance to features in the real world (typically in the field of computer vision and tasks pertaining to that research field), the so-called why-questions can be tackled as well. Zednik here relies on the example of models deployed for recognition of armored fighting vehicles: ANN trained for distinguishing, say, a T-14 Armata from a T-84 Oplot may do so by focusing on accidental association based on environmental features. Input heatmapping could help trace such features, which would, in turn, inform examiners at the International Criminal Court or decision subjects in army units or military academies. To sum up, techniques such as input heatmapping can help us to discern genuine regularities (be it causal or statistical) in ANNs from artifacts. In this way, the ML or DL model becomes interpretable regarding interconnections between input and output.

Boge's considerations seem to overestimate the quality and quantity of epistemically relevant elements needed for examiner, data, and decision subjects. This is because he does not go into detail regarding diverse types of agents seeking answers, but, as we have seen in the previous passage, the devil is precisely in the details. What remains to be seen is whether how-questions and epistemic status of creators can be boosted to avoid Boge's pessimism since this particular case is best aligned with his stronger version of the black-box problem regarding the inability to explain the model's behavior in principle. I will advance my argumentation by drawing on Zednik & Boelsen (2022), who argue that XAI models should be viewed as *exploratory* — as means for generating and testing hypotheses and refining target phenomena. This proposal seems particularly suitable for two reasons:

[51] A brief *nota bene* would be helpful here: in linear models, each feature is equally important for the output because it has the same value for each data point in the input layer; albeit in non-linear models the values diverge for each data point, and it is paramount to find means to calculate such values to make non-linear models interpretable. Input heatmapping is an umbrella term for a family of XAI techniques, such as *Layerwise Relevance Propagation* (Montavon et al. 2018) and *Shappley Additive Explanation* (Lundberg & Lee 2017). The main job of this family of XAI techniques is to average out the different values of regions where the dataset is being transformed into the input and then distributed through an ANN. Thus, the first technique is particularly suitable for the research on computer vision since the whole point is to highlight relevant pixel regions in images fed to the input layer that are statistically correlated to output. The second technique is more widely used since it is based on using Shapley values to calculate the impact of all features of the output since more salient features usually have larger weights and, therefore, their impact is more relevant when validating the model's behavior.

(*I*) Given that the XAI approach presumes constructing internal models and employing post-hoc techniques through which it is possible to make ML models interpretable, and I have already argued that DL models in cognitive science are best understood as exploratory, it seems fairly sensible to extend this point to their internal models as well.

(*II*) Zednik & Boelsen emphasize that this exploratory role should be spelled out in comparison to cognitive science, in a similar way as Zednik (2019) used Marr's three levels as a heuristic to develop a normative framework for the XAI approach. In this way, a close link between XAI and cognitive science is forged, reinforcing the other reflections in this dissertation about the feasibility of the connectionist cognitive architecture.

Now, to see why Boge's worries should not overwhelm us, it is enough to bring all the pieces together: if some XAI internal models have an exploratory role manifested through specifying and assessing target phenomena for suitability, as well as generating and testing hypotheses about phenomena, then the upshot of XAI is to *chart the search space of possible answers to how-questions (*as conceived by Zednik). Rather, by searching for answers to how-questions pertaining to internal mechanisms, a modeler is trying to make her DL model interpretable in sense that she wants to shed light on the manner in which input and output layers can be interconnected and thus explain why the output of her model should be taken at face value by a wider (be it expert or layperson) audience. As I have already mentioned, exploratory models are frequently invoked when we want to *refute an impossibility claim* (Sjölin Wirling & Grüne-Yanoff 2021). Thus, it is convenient that this is precisely what we are aiming for through XAI approach since the goal is to explore possible inner mechanism of a particular DL model or to refute the impossibility claim pertaining to its explanatory ambition. In this sense, to claim that XAI techniques are in principle unable to answer *how-questions* amounts to assuming the conclusion.

Even if one grants that I have discarded Boge's pessimism, the following question remains unanswered: what sorts of explanation can we get from analyzing post-connectionist models, though? I follow Stinson (2018, 2020), along with classics such as Machamer, Darden, and Craver (2000), Craver (2007), and Darden (2008) in adopting the mechanistic account of explanation. This account presumes three distinct stages of describing a mechanism, viz., through a sketch, schema, or complete and detailed description. At the first stage, constituent parts of a mechanism are known, at the second stage, we are familiar with parts albeit not with their interaction, and, finally, at the third stage we have all the relevant details pertaining to the organization of the mechanism. As Bechtel & Richardson (2000) claim, mechanistic explanations have two distinctive features that make them suitable for cognitive and neuroscience: (*i*) they explain in virtue of referring to *functions* that each constitutive part of a system performs, and (*ii*) they are obtained through the process of *decomposition* of a system to its constitutive parts and their interactions. Connectionist models function similarly to (*i*): in order to find answers to *what*- and *why*-questions in Zednik's terminology, we ought to understand the functions of units, parameters, and hyperparameters within a specific computational architecture. More difficult questions, viz., *how*- and *where*-questions, are to be tackled through decomposition which can be done thanks to post-hoc analytic techniques. What unites these questions is the conviction that all can be answered by relying on the shared abstract mechanistic structure that justifies drawing inferences from connectionist models to target systems.

Thus, DL models in cognitive and neuroscience, like connectionist models in general, offer mechanistic explanations thanks to constraints and details stemming from physiology and anatomy, which constitute abstract mechanistic structure that is shared with brain as per functionalist assumption. First, such constraints—embedded in different architectural features and hyperparameters—serve as inferential pincer movements that help narrow down the space of possible cognitive mechanisms (Stinson 2018). This is in line with the exploratory role of DL models. Second, physiological and anatomical details serve for probing the mechanism design within the model. The last step is to validate insights regarding the interaction and organization of the mechanism parts in an actual lab—via independent data such as fMRI, results from experimental studies in psychology, etc. The three steps align well with three stages of mechanism discovery through the explanation, i.e., sketch, schema, and complete description of a mechanism. Additionally, the three steps can be easily related to the point I made earlier about connectionism being an undeveloped research program and a part of the immature scientific field: the explanatory dynamics of mechanism discovery is to be expected when the research program still develops tools and *explananda* for looking into *explanans*.

The progress from focusing on constraints, then on adding details, and finally on validation is similar to Craver's (2007) distinction between how-possibly, how-plausibly and how-actually explanations. How-possibly explanations are most tightly intertwined with the exploratory role of both DL models and post-analytic techniques that should shed light on their inner functioning. DL models that provide us with such explanations are highly idealized and sketchy—usually with no or only a few slightly biologically plausible constraints. These explanations, as I have already hinted, provide proof of principle demonstrations, i.e., what could be done within a model, or what possible or impossible *explananda* can or cannot be tackled within a model. *Explanans* can also be within the domain of possible and impossible. To put it in Zednik & Boelsen's words, XAI approach can, thus, "facilitate the specification of algorithms to be considered as possible explanations of behavioral or cognitive phenomena" (2022: 232). Most often, how-possibly explanations have to do with technological innovation needed to surpass previous methodological impasses in connectionist modelling. Thus, for instance, Elman (1990, 1991) introduced SRN as means to explore whether connectionist models can be used *at all* for higher cognitive processes with an open-ended nature, such as natural language, given the bad reputation of connectionism (recall **2.3.**). In other words, Elman wanted to see whether such computational architecture along with its novel features could provide us with a possible explanation of human processing of grammatical structure.

How-plausibly explanations arise when there is a tradeoff between biological plausibility of a model and its computational or cognitive efficacity within a DL model. The upshot is to add relevant details but not at the expense of computational efficacity. Whereas how-possibly explanations can be obtained from sketchy DL models, how-plausibly explanations can be obtained from schematic DL models, in which architectural features are specified, along with properties of target system to be modeled, but further details about how specific behavioral mimicry of a model emerges are casted about. Thus, a modeler focuses on whether any crucial property has been left out or altered by watching closely the qualitative changes in a model. In this way, she can see what happens to the mechanism once some parts are changed or broken down, which, in turn, helps tracing the interaction between the parts, their overall organization and their alleged contribution to certain

behavioral patterns. AlexNet starts with biologically plausible architectural features entrenched in Hubel & Wiesel's research of feline visual cortex and is, therefore, very specific about properties of the target system. However, the task of AlexNet is to offer efficient classification of images. A modeler must inspect what features are relevant for such a task and what interactions within a model result in successful task performance. The more physiological details pertaining to vision researchers add, and the more plausible DL models based on DCNNs become, i.e., they start to represent human visual processing more faithfully. I do not want to give a false impression here that how-plausibly explanations are tied exclusively to DCNNs. Instead, when it comes to psychological plausibility, this sort of mechanistic explanation is applicable to discussions about LSTMs, RNNs and transformers in the context of NLP. Thus, cognitive scientists and psychologists, Brendan Lake and Gregory Murphy, examine whether NLP models can serve as accounts of psychological semantics and note that what matters is "[t]he question [w]hether the model's processes are *plausibly* similar to those of humans, possibly giving insight into human psychology" (2021: 25, my emphasis).

Buckner (2023) would go further and claim that DL models can offer us explanations of how humans obtain knowledge about the world. I would, however, stress that this is nowhere near as direct and linear a process as Buckner would want it to be. Rather, the exploratory role of both DL models and XAI approach that serves to render them more transparent show the explanatory dynamics of gradual mechanism discovery in cognitive science and the slow but steady constitution of connectionism as a more mature research program. The explanatory dynamics emerges in cases that Uljana Feest (2017: 1166) describes as "[c]ases when *explanantia* and *explanandum* are subject to the ongoing process of investigation" and "*explanantia*, must be precisely characterized" while "*[e]xplanandum* is constantly changing." The leap from how-possibly to how-plausibly explanations is dependent on epistemic goals of modelers and independent evidence stemming from empirical studies such as psychology or neuroscience that help calibrate DL models. Along the way, as models become technically refined and (more) transparent, how-actually explanations could be advanced albeit only in the domains where *explanantia* is precisely characterized, i.e., where we grasped the complete mechanism of both model and target-system *and* have relevant amount of independent evidence. I applaud Buckner's optimism that represents a stark contrast to Boge's pessimism. Nonetheless, the only candidates for offering how-actually explanations in the near future are, in my view, DCNNs and their application in the cognitive neuroscience of vision. This is not at all controversial claim, given that neuroscientists extensively use DL models implementing DCNNs as tools for probing specific hypotheses about object recognition (for an overview see Lindsay 2021) and these models have been validated by the other neuroscientific methods, say human fMRI as in (Güçlü & van Gerven 2015), or representation similarity analysis as in (Khaligh-Razavi & Kriegeskorte 2014). Furthermore, this is not surprising either; connectionist models from their inception have quite accurately captured lower cognitive processes such as perception (recall Sect. **2.3.**). Optimism regarding the feasibility of how-actually explanations in other domains and for other cognitive processes is yet to be empirically justified. One promising step, however, can be embodiment, i.e., incorporation of ideas from embodied approaches to cognition in multimodal post-connectionist models. I will return to this point in Sect. **4.3.**

To sum up, the main aim of this gruesomely technical Sect. was to cover methodological intricacies of the current post-connectionist models to spell their epistemic role and general explanatory prospects given the most serious methodological issue, viz., the black box problem. I have argued that this problem is not insurmountable in the case of cognitive processes: DL models offer us mechanistic explanations of cognitive processes which may differ with respect to the level of grain that could be currently obtained. I hope that I have convinced the readers that these models are not "more of the same", as the criticism goes, but rather represent a genuine methodological breakthrough in AI and fresh meat for linguistics, cognitive science, and neuroscience. Alas, this is nowhere near establishing that they are not "more of the same" regarding their empiricist ambitions. This is a task for the next gruesomely philosophical Sect.

3.2. Deep Learning: Failed Ambitions or the Startling Advantage of Neo-Empiricism?

So far, I have argued for the following points:

(A) Post-connectionist models are *not* threatened by their alleged biological implausibility, nor are they inferior in comparison to the shallow connectionist models due to the alleged biological implausibility. Their role is rather exploratory. Modelers get to tinker with the parts and see how they interact, thereby discovering generic mechanisms that underly cognitive processes.

(B) Post-connectionist models are *not* threatened by the black box problem either. They are explanatory in a sense that they should be understood as mechanism sketches that are gradually being filled in with relevant details from cognitive and neuroscience so that we could proceed from how-possibly to how-plausibly stage of understanding cognitive mechanisms. During this process, these models are calibrated, and their empirical adequacy is being re-established despite the fact that they are representationally inaccurate (or not as accurate as philosophers would want them to be). XAI approach—envisaged for rendering DL models transparent—reveals inner mechanisms of DL models in a parallel manner, i.e., by exploring the space of possible and plausible inner workings of a particular DL model.

However, the time has come to evaluate much more worrisome charges coming from the nativists. History repeats itself once again. Basically, the same criticism that was advanced against shallow connectionist models is now invoked against post-connectionist models regardless of their methodological differences (recall **Tab. 3**). The general structure of the heated argument exchange between nativists and empiricists looks roughly like this (Milojević & Subotić 2020: 136-137):

(1) A specific connectionist architecture or model is proposed.
(2) Nativists claim that (1) cannot account for a cognitive phenomenon Φ due to the inability of (1) to *structurally* subtend mechanisms needed for Φ.
(3) Empiricists, in turn, claim either that:

a. The model does not have to explain Φ per se, but rather the implementation of Φ, i.e., the lower (subsymbolic or subpersonal) level where Φ is realized.
b. The model can be structurally changed through technological innovation so that criteria for subtending Φ can, in fact, be met in due time. This results in (1′), i.e., a novel architecture or model.

(4) Nativists discard both (3a) & (3b) by maintaining that:

a. Connectionist models fail in the implementation of Φ since they are not biologically plausible.
b. There is a novel phenomenon Φ' that cannot be explained by (1′) which resulted from (3b). This further entails that connectionism is in principle unable to offer adequate explanations of human cognitive processes.

Essentially, my task here is to evaluate the brand-new instance of the old empiricism vs. rationalism debate across AI research, cognitive science, and linguistics, so that I could defend the banners of moderate neo-empiricist dogma in the domain of NLP in the next Sect. Important thing to note here is that the possibly worst-case scenario—namely, that LLMs do not offer us an adequate empiricist account of linguistic competence—could not undermine the general neo-empiricist dogma that is at stake in this Sect. In other words, my task here is to provide a framework that safeguards **Specific hypotheses II** and **III**. Recall, **Specific hypothesis II** establishes the difference in labeling between contemporary philosophers and scientists and those of previous eras, i.e., being an empiricist or rationalist nowadays comes with additional theoretical commitments; while **Specific hypothesis III** aspires to set out connectionism as an autonomous theory of human cognition rather than mere unorthodox approach to computational modeling. Of course, the worst-case scenario would go against the **Main hypothesis** that *can* be defended, but as I have indicated in the **Introduction**, I would be perfectly content if the reader acknowledges at least some of the incremental steps in the overall argumentation within this dissertation. Specific hypotheses are, in any case, independent of the main and auxiliary hypotheses. The maneuver to safeguard specific hypotheses hinges on (*A*) and (*B*). Once the exploratory role of post-connectionist models is acknowledged, along with the explanatory dynamics of gradual progression from how-possibly to how-plausibly and how-actually explanations, these models become immune to attacks like (2), or (4a) & (4b). First, however, I will present the strongest rationalist and empiricist voices within the current debate. Let the games begin.

A Strawperson Empiricist and Impartial Rationalist Enter a Bar...

Labeling in any scientific field, but especially in philosophy, is a tricky thing indeed. Labels help us differentiate between "us" and "them", between "our" credo and their "gibberish". When labeling happens across several fields, then a "tricky thing" becomes an understatement. Many AI researchers have seen themselves as either "empiricists" or "rationalists" from the inception of computer science. Some of them have been "enlightened" by philosophical discussions, some of them were rather appalled by such discussions, and some of them were just using buzzwords. The same goes for cognitive scientists. The mission here is to disclose the true believers from the flaunters, which essentially amounts to detecting "mock" rationalism and "mock" empiricism, i.e., radical and theoretically unrefined grand claims that are of no use for either advancing the debate

or understanding what is at stake in the debate. These claims often stem from the habit of contemporary scientists to conflate (E_{CS}) and (R_{CS}) with (E_H) and (R_H).[52] Their claims are usually full of historical terminology ("*tabula rasa*", "innate endowment", etc.) and their understanding of the labels they are choosing for themselves is more superficial than based on a previous inquiry about labels' connotations. Quite the contrary, historical terminology is completely inadequate in this regard since their views can only be similar to some extent to (E_H) and (R_H), but with the additional gloss in the form of domain-specific and domain-general distinction. This is what I will call "moderate rationalism" and "moderate empiricism", which are, at the same time, only acceptable positions if the whole point of the debate is to deepen knowledge about the origins and nature of cognitive processes, including NLP.

As in Ch. **1** & **2**, I am concerned with *psychological nativism* (Cowie 1999) or *origin rationalism* (Buckner 2023), and I use the terms "rationalism" and "nativism" interchangeably. Contrary to Ch. **1**, the rationalism vs. empiricism debate should no longer be seen as an *a priori* debate about knowledge. Contrary to Ch. **2**, the rationalism vs. empiricism debate should no longer be seen as a modeling or engineering competition— the success of DL models is unprecedented in the history of computer science (recall Sect. **3.1.**). Instead, I will make my case that the debate cuts across the foundations of cognitive science, i.e., the very methodology of understanding human cognition through computational models and the core assumption that structural perks of a computational architecture mirror those of a cognitive architecture. This brings the defense and confirmation of the specific hypothesis **II** to its conclusion.

Throughout the dissertation, I have introduced the difference between supervised and unsupervised learning. Unsupervised learning is the closest engineering feature to the empiricist idea of gradual filling-in of the *tabula rasa* through the sensory-motor interaction

[52] To refresh your memory, I rehearse here what is assumed by (E_{CS}), (R_{CS}), (E_H), and (R_H). (E_{CS}) states that the cognitive processes are best simulated and examined by computational models implementing ANNs. These models are designed to minimize the manual specification of rules and to rely on learning algorithms to prompt ANNs to learn from experience, i.e., data. On the other hand, (R_{CS}) states that the cognitive processes are best simulated and examined by computational models implementing symbolic representations and rules for manipulation over representations. These models are designed to produce human-matching task performance based on the list of instructions that encode expert knowledge. Both (E_{CS}) and (R_{CS}) thus reflect competing scientific understandings of the relationship between computational and cognitive architecture which may or may not have philosophical underpinning. This is an engineering rivalry: what computational architectures, viz., what type of models would live up to be the main protagonist of the Cinderella story and reveal the nature of our cognition at midnight. However, (E_H) holds that the contents of human minds, i.e., knowledge, and capacities constituting the human minds are grounded in sensory experience and to the great extent acquired rather than innate. Experience is also a touchstone of meaning, truth, and/or any abstract notion whatsoever. The adversarial position (R_H) maintains that he contents of human minds, i.e., knowledge, and capacities constituting the human minds are shaped by our innate rationality with which God endowed us and thus made us unique in the nature. This is also a touchstone of meaning, truth, and/or any abstract notion whatsoever. Both (E_H) and (R_H) mark a specific period in the history of philosophy which gave birth to the conceptual rather than engineering rivalry that will shape the philosophical and scientific pursuits of centuries to come. However, the terminology used by proponents of the two positions is imbued with philosophical and metaphysical legacy that need not have counterpart in cognitive science or AI research, and indeed, it rarely has unless one makes the mistake of using the terminology regardless of its historical context.

with the environment. Thus, David Silver and his colleagues start their paper with the following statement:

"A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains [...] Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo." (2017: 354)

AlphaGo Zero, an offspring of the *Google DeepMind*'s laboratory, was trained in an unsupervised manner, through deep reinforcement learning, and without relying on the expert data sets. In other words, the model was more advanced than the previous iteration (dubbed "AlphaGo Lee") that managed to outsmart Lee "The Strong Stone" Sedol, one of the best Go players in the world holding 18 international titles. Silver and colleagues maintain that this remarkable success is due to training through self-play:

"AlphaGo Zero outperformed AlphaGo Lee after just 36h. [...] After 72h, we evaluated AlphaGo Zero against the exact version of AlphaGo Lee that defeated Lee Sedol, under the same 2h time controls and match conditions that were used in the man– machine match in Seoul. [...] AlphaGo Zero defeated AlphaGo Lee by 100 games to 0." (2017: 356)

Finally, not only did AlphaGo Zero master human knowledge of Go even though it started with random moves, but the model also exhibited non-standard strategic moves, that go far beyond human knowledge of Go. This suggests some sort of domain-limited creativity in a model, perhaps rudimentary capability for abstraction as well. Humankind played Go—an abstract strategy board game whose number of possible matches exceeds the number 10^{768}—for more than 2500 years (*American Go Association* 2017), and yet, a post-connectionist model implementing a DCNN, such as AlphaGo Zero, crunched such time period into a few days and produced novel expert patterns. Humanity 0: AI 1. Also, rationalism 0: empiricism 1. [53]

However, if things were so simple, this dissertation would have been much shorter. Gary Marcus, a cognitive scientist and AI researcher, has kept an eye on connectionism since the 1990s given that he was and still is a fervent supporter of the marriage between traditional symbolic science and GOFAI (see Marcus 2001). Furthermore, his support hinges on empirical work in developmental psychology. Thus, Marcus (2018b: 3) starts with the premise that evidence of prewiring does not speak against rewiring, or that evidence for rewiring speaks against prewiring. This essentially means that studies about *core cognition*, i.e., knowledge of naïve physics (Spelke 1994, Spelke & Kinzler 2007), or about 8-month-old infants learning abstract rules from a short exposure to artificial grammar (Gervain, Berent & Werker 2012), suggests that we come to this world with an innate armamentarium which makes us efficient cognizers unlike data-hungry and computationally costly DL models, such as AlphaGo Zero. This is not to say that such models cannot capture something about

[53] Another fervent radical empiricist is Yann LeCun. However, he did not express his allegiance in published papers, but rather in public, during debates such as this one from 2017 with Gary Marcus and David Chalmers as moderator. The crux of the debate was whether DL models need more innate machinery and, as expected, LeCun advocated for radical empiricism, whereas Marcus fluctuated between radical and moderate nativism. I return to this debate in the next Subsect.

us, i.e., our rewiring through learning to adapt, but they cannot simulate our way of *cognizing*.

Marcus takes cognition to be a function of four variables, namely innate algorithms (*a*), representational format (*r*), domain-specific or domain-general innate knowledge (*k*), and experience (*e*). Radical empiricists would want (*a*), (*r*), and (*k*) to approach zero, and (*e*) to approach 1, while nativists would want the opposite.[54] However, no post-connectionist model actually looks like the radical empiricists' version, not even AlphaGo Zero.[55] As Marcus remarks, although AlphaGo Zero generates its own training databases, the model does have priors in the form of game-specific representations and elementary instructions, thus not being truly faithful to unsupervised learning. For instance, the DCNN implemented in a model included a special algorithm for handling Go board in realistic circumstances, i.e., reflections and rotations of the board, instead of extracting the right behavior from mere visual input. Thus, instead of being a landmark model of radical empiricism, "AlphaGo is actually an illustration of the opposite: of the power of building in the right stuff to begin with" (Marcus 2018b: 8). Finally, Marcus jumps from remarking that game-oriented models cannot work properly without hardwired tree structures as search heuristics, such as Monte Carlo tree search to concluding that—as per nativist admonition—*all* DL models must incorporate such structures for *any* cognitive task. Recall the general structure of the argument exchange between empiricists and nativists. Marcus is not (much) concerned with the biological plausibility of AlphaGo Zero, but rather with its structural inadequacies. He also does not make us wait long for his ultimate assessment of any post-connectionist model that has been or will be created, thereby kindly shortening the argument exchange. Humanity 1: AI 0. Rationalism $+\infty$: empiricism $-\infty$. Or, to put it in his notation: if the model purports to simulate human cognition, the values of (*a*), (*r*), and (*k*) must approach 1, while the value of (*e*) is mainly irrelevant. Computational architecture without tree structures fails to account for cognitive architecture that—*ex hypothesi*—must contain innate machinery akin to tree structures, ergo post-connectionism presents us with wrong methodology for understanding human cognition or inferring anything about human cognition from it.

Jest aside, Marcus has also published a separate piece that deals with the shortcomings of DL generally (2018a). There he offers a broader perspective on his pessimism about empiricist ambitions embodied in a single game-oriented model, such as

---

[54] Bear in mind that what Marcus simply dubs (*a*) and (*k*) actually reflects the difference between inductive biases and priors. The former guides the learning process and amounts to predefined computational constraints, the latter represents initial distribution of probabilities that are based on the encoded knowledge of task performance. Arguably, AI engineers with empiricist ambitions are not fooling themselves that there are no inductive biases in the model, whether it be in the case of supervised or unsupervised learning. Check Goyal & Bengio (2020) for a full list of inductive biases that they consider indispensable for DL models in order to simulate higher-level cognition.

[55] Moreover, the so-called "No Free Lunch Theorem" (Walpert & Macready 1997) shows that any (*a*) that works well for a particular set of problems, pays for the optimal performance with poor performance for the other set of problems, and that, strictly speaking, (*a*) cannot be literally zero. Hence, one could argue that there are mathematical reasons why radical empiricism is unfeasible in computer science. I will return to the consequences of this theorem later on.

AlphaGo Zero. Marcus echoes the same shortcomings that Fodor & Pylyshyn (1988) advanced against the "shallow" connectionism. The sole difference is what exactly is innate—rules or hierarchical structures, albeit this difference is grounded in semantics not ontology. Additionally, Marcus sums up other "fresher" albeit frequent disapprovals of DL, such as models being too data-hungry, limited in terms of knowledge transfer, and not being sufficiently transparent (see 2018a: 6-12). The very method of undermining DL consists of taking these four key issues and creating several piecemeal issues out of the original ones, thereby creating an impression that any proponent of DL cannot possibly begin with waging defense in light of myriads of insurmountable problems. For instance, the black box problem, or the lack of transparency, is presented as a problem on its own, but a few pages later we can read about a separate issue of DL not being able to distinguish causation from correlation, or that it cannot be trusted, which are all either corollaries or consequences of the original problem. I will not engage in tackling his criticism of DL since I have already done that in the previous Sect. My goal here is to debunk nativist commitments and argumentative strategies embedded in such criticism. However, Marcus also backs off at some point and admits that cognitive neuroscience and developmental psychology may lend a hand in making DL models more robust, comprehensive, and faithful to human cognitive processing. Moreover, in line with Buckner (2023), he proposes to use DL as a tool for probing our currently best hypotheses about inner mechanisms, which would be beneficial for understanding both minds and machines. This may strike a reader as odd, as did me, given that the previously advanced knockout strategy did not really leave any space for admitting that DL, as it currently stands, does suggest that connectionist cognitive architecture can be regarded as a worthy opponent to the symbolic one—at least, worthy enough to be probed and boosted with details from cognitive neuroscience. This is also something which I have defended in the previous Sect.

The main issue with the knockout nativists' strategy is that it trivializes the debate and can be used against them in the same manner. Let me show how this strategy can be used from radical empiricists perspective as well. DeepBlue, a chess-playing expert system based on GOFAI, won a rematch with Gary Kasparov in 1997. It was an inspiring feat that allegedly attested how implemented rationalist agenda mirrors our cognitive efficiency embedded in hardwired instructions within the DeepBlue's symbolic "mind". The alternative would be learning from every single match. However, unfortunately for DeepBlue, it had coarse-grained representations applicable to narrow domain, namely chess playing, that could not account for flexible cognitive performance. In other words, its computational structure is inapt for subtending mechanisms underlying creative, intuitive, or strategic board games. Furthermore, any other cognitive task would be simulated in an overly brittle manner, thus *any* model based on GOFAI, no matter how advanced or refined, is doomed precisely because it cannot learn from experience. Human expertise is constituted from both knowledge-that (i.e., instructions) and knowledge-how (i.e., skills), and this mixture further enforces cognitive flexibility in previously unseen or unexperienced circumstances (for an overview of empirical studies of expertise see Ericsson et al. 2018, for historically important theoretical pieces, see Ryle 1949 and Polanyi 1966, see Chase & Simon 1973 for chess players' expertise). Lucky for us, we have models based on biologically plausible ANNs that do learn from experience, i.e., the chunks of numerically

transformed real-world data, thanks to their structural peculiarities, such as different computational architectures trained through a variety of learning algorithms. Empiricism +∞: rationalism -∞. There are, nonetheless, many problems with the empiricist knockout strategy, which are virtually the same as with Marcus (2018a, 2018b). First, it is materially false. Hybrid models, that I will be tackling in the next Ch., excel in many tasks that were too demanding for classic expert systems. The same goes for a future-oriented research program such as connectionism, which produces models with an exploratory role, including AlphaGo Zero. If anything, we have seen that AI researchers have pushed forward post-connectionist models with respect to the quality and quantity of their cognitive task performance, and I have described this progress in terms of the progress from how-possibly to (at least) how-plausibly and (hopefully) how-actually explanations. Second, the all-or-nothing claims inferred from the performance of a single model have no argumentative use but to halt any constructive discussion and obtain media coverage to the one who uttered them. Third, we have not even scratched the surface of the crucial issue regarding the inner workings of a model, i.e., whether (*a*) and (*k*) are domain-specific or domain-general, and their relation to *(r)* as per our professed position.

The moral of the AlphaGo clash and DeepBlue counterexample is that one side, usually professed to be impartial and more scientifically rigorous, ends up attacking the strawperson with sticks and stones. This is more frequent than reader would guess, and interestingly enough, more typical of contemporary exchanges between connectionists and symbolists than it was the case in the 1980s. I believe that one of the reasons for this can be traced to the general detachment of the current post-connectionist models from theoretical frameworks and interdisciplinary efforts due to their lucrative industrial usage that I have mentioned at the beginning of this Ch. The laboratories within tech giants such as Google, Microsoft, or Meta are allowed to push the boundaries of computer science as far as profit goes. This results in a less refined peer disagreement since at the back of one's head are press covers, and press is full of good guys vs. bad guys clichés (after all, this is another all-too-familiar moment in the history of connectionism, recall Sect. **2.3.**). Moreover, it turns out that in virtue of computer and cognitive scientists' labels, no contemporary philosopher has ever been radical empiricist or nativists.

To see how even relatively independent researchers are not far from this unflattering image of peer disagreement, take another example into account. Judea Pearl, a computer scientist and philosopher working on causal and counterfactual inference in machines, also targets a position that is much like "mock" empiricism rather than actually avowed empiricism in his recent technical report. However, unlike Marcus, who is far from a "mock" nativist, Pearl blackens the name of their professed position while trying to debase "mock" empiricism. Pearl uses the distinction *model-based* vs. *model-free* approach to account for the difference between supervised and unsupervised learning. Model-based approach assumes that engineers start with the model of the world, i.e., obtain as much as possible input from the modeled system and fine-tune parameters in such a way that the model learns and builds its responses to stimuli as realistically as possible (Lake et al. 2017: 3). On the other hand, the model-free approach assumes that engineers start with the predictions of model's behavior by letting it find on its own shared values during learning or shared

features during classification tasks (Lake et al. 2017: 2-3). Pearl sees this difference in evolutionary light. Those favoring model-free approach are prompted by "data-centric" Western philosophy because they believe that

> "in order to develop human level intelligence, we should merely trace the way our ancestors did it, and simulate both genetic and cultural evolutions on a digital machine, taking as input all the data we can possibly collect" (Pearl 2021: 78).

Here, taking as input all collected data would be akin to allegedly empiricist agenda of promoting sensory experience as the sole source of "all our concepts and knowledge, with little or no role given to 'innate ideas' and 'reason' as sources of knowledge" (Pearl 2021: 79). Furthermore, Pearl explicitly equates connectionism with radical empiricism and tags Cameron Buckner as a typical radical empiricist.[56]

Pearl's completely distorted view of the position that he is criticizing reveals a mash of logical fallacies, namely the fallacy of composition and *petitio principii* in disguise. He concludes that connectionism is a radical empiricist position by assuming that it relies on unsupervised learning techniques, which mimic traditional empiricist *tabula rasa* image of cognition. Of course, Pearl can still think that he has a good point there, given the hubristic claims of Silver et al. (2017). However, there is a significant difference between claiming that X is Y because Y is one of the constitutive parts of X, which assimilates X altogether, and inconsistent labeling in Silver et al. (2017). As we have seen, Silver and colleagues reported specific biases and, hence, they indirectly admitted fine-tuning of the model despite the grand claims about AlphaGo Zero being an exemplar of *tabula rasa*. Pearl presents a mock empiricist position by fixating radical empiricist notion of *tabula rasa* in order to argue for the inclusion of "tools and principles of causal science to guide data exploration and data interpretation processes" (2021: 80). In other words, he builds his case against DL (and ML in general) by pointing out that engineers must incorporate into post-connectionist models means for running counterfactual and hypothetical theories thanks to the built-in world models, which essentially means that Pearl calls for more innate machinery[57] in such models, as every other nativist does. Unlike Marcus, however, Pearl does not give the benefit of a doubt to his opponents.

Finally, I owe the reader the example of true believers who defend banners of their positions without adhering to either of the two "mock" strategies. This could be due to the fact that unlike AI researchers and cognitive scientists, the true believers are usually philosophers or scientists with a strong background in philosophy. Hence, the true

[56] Buckner's paper cited in Pearl (2021) is, in fact, a review paper that surveys various aspects of deep learning relevant for philosophers of mind and philosophers of science. Not a single sentence in the paper gives the impression that Buckner endorses radical empiricism, or any position whatsoever, given that it is a review paper. In his other publications, Buckner is explicit about being moderate empiricist (see Buckner 2018, Buckner 2023), and I will examine and further develop his position in the rest of this Sect. Nonetheless, it is worth noting here that bad scholarship, which favors ideology and confirmation of already established unipolar opinions over a genuine exchange of arguments, had already contributed to the first AI winter when Minsky and Papert published their *Perceptrons* (recall Sect. **2.3.**).

[57] Truth be told, it is not easy to discern whether Pearl considers causal and counterfactual inferences as innate knowledge or innate mechanisms within a DL model.

believers often want to entrench their claims in the history of philosophy to establish a philosophical link between engineering and theoretical aspects of DL by invoking historical discussions as means for legitimizing such endeavor. Interestingly enough, the entrenching of the claims only differs in how deep the authors are willing to go. The empiricist banners are defended by (who else than) Buckner (2018) through drawing on British empiricism from early modern period (Sect. **1.3.**), while Childers, Hvorecky, and Majer (2021) defend rationalist banners through waging an attack on analytic empiricism (Sect. **1.4.**). As we shall see, both parties rehearse some of the points made by their bolder peers, albeit in a more convincing manner.

Childers and colleagues begin their paper with a strong and contentious claim that Skinner's behaviorism had an overwhelming methodological influence on connectionism and contemporary DL models. The key behaviorist idea that is used for backing up this claim is the following: "The strategy envisioned by the behaviorist theory of the mind aims at a creation of an implicit list of probabilistic correlations of stimuli and responses—and little else" (Childers, Hvorecky & Majer 2021: 2). Basically, this is how any connectionist model works—a list of probabilistic correlations among many activated units that results in a distributed representation and emergent behavior of an ANN. Among philosophers, W. V. O. Quine endorsed Skinner's behaviorism (e.g., explicitly in 1960/2013) and through his dispute with Chomsky (1969), as well as Chomsky's influential criticism of Skinner (1959), Childers and colleagues trace the dispute between connectionists and symbolists, i.e., proponents of DL and proponents of GOFAI. I have already tackled this dispute in Sect. **2.1.**, therefore, I will not rehearse the details. The important thing here is the reinterpretation of it in terms of DL vs. GOFAI. The authors see this dispute as the first of many hybridizations that empiricists are obliged to undertake when faced with rationalist criticism. In other words, they start enriching their originally "pure" picture with representational elements. This is virtually the same claim that can be found in Marcus (2018a, 2018b). Also, despite connectionism being explicitly committed to representationalism, Childers and colleagues again turn to the reinterpretation and regard the history of connectionism in terms of the coveted triumph of anti-representationalism, which is, essentially, behaviorism:

> "Psychological terms and their adoption at the level of scientific mentalism would have to go, and the doors of empiricism would remain wide open" (2021: 8).

Instead of focusing on unsupervised learning, like Pearl (2021), these authors cover both supervised and unsupervised learning, albeit in an asymmetrical manner. When it comes to the supervised learning, they remark that the input of an ANN is always pre-processed by a human modeler or subject, therefore, *sensu stricto*, no input is raw input. Rather, when a model starts to process input on its own, its intensions are different than ours: we pick different features, create different categories, have distinctive semantical vehicles of content. That is, if we can ascribe intensions to ANNs at all, given that they operate on stochastic grounds. Thus, there are three steps of argumentation in Childers, Hvorecky, and Majer (2021):

(1) Any sort of fine-tuning in connectionist and post-connectionist models entails the betraying of empiricist agenda, and, *ipso facto*, behaviorist agenda.
(2) There is always some sort of fine-tuning present in connectionist and post-connectionist models, at least in the case of supervised learning.[58]
(3) Even when put to work, processing in connectionist and post-connectionist models is incommensurable to our cognitive processing: intensions cannot be reduced to extensions, i.e., mere frequencies of pixels (DCNN-based models) or words (transformer-based or RNN-based models).

When it comes to the unsupervised learning, the authors pass the burden of proof to Cameron Buckner, again labeled as a radical empiricist believing that he can defend such an absurd position through unsupervised learning, when we know that every model has some parameters and hyperparameters that are hardwired. In other words, every model, regardless of the choice of training algorithm, comes with priors. Now, recall (1) & (2). Check mate, right? It is important to note here that, unlike Marcus and Pearl, Childers and colleagues are waging a substantial philosophical attack against empiricist ambitions of DL proponents as (3) shows. They are stressing the ontological difference in inner workings of the human mind with respect to post-connectionist models instead of focusing only on the methodological limitations of specific models and concluding from it that connectionism fails as a hypothesis about human cognitive architecture. By claiming that this difference is grounded in intensions, i.e., our ability to understand the meaning of words, instructions, and our environment, they essentially sugarcoat the assumption of human exceptionalism coupled with rationalism of the 20th-century cognitive scientists and linguists. Hence, we have all elements needed for taking the debate one step further: historical trajectory that reveals key theoretical assumptions behind methodological decisions of modelers and metaphysical commitments as paraphernalia for intensifying the discussion.

Cameron Buckner, so far notorious for his "radical" empiricism, in one of the first papers tackling philosophical aspects of DL (Buckner 2018), aims to show how DL can help us shed light on the higher cognitive processes, such as abstraction. This process is both cognitively and philosophically interesting because British empiricists had a lot to say about it (recall Sect. **1.2.**). Moreover, even though simulating abstraction would not be enough to show that ANNs have the potential to constitute the underlying architecture of the artificial general intelligence, it would be valuable evidence in favor of the viability of connectionist cognitive architecture. As we have seen in Sect. **2.3.**, the crucial issue of "shallow" connectionism was its inability to successfully account for higher cognitive processes in the same way as it could for lower cognitive processes. Additionally, Buckner's argument in favor of DL models hinges on the specific type of ANNs, namely DCNNs. If victorious, his defense of empiricist banners through DCNNs when it comes to abstraction may serve as both a source of inspiration and resource for the defense of my main hypothesis.

Simply put, Buckner thinks that DCNNs can and do model the process of abstracting exemplars from experience and creating new ones. The type of abstraction he has in mind is categorical abstraction, i.e., the manner in which one represents category membership.

[58] The same goes for semi-supervised learning as in the case of, say, ChatGPT.

However, the main thing Buckner tackles is how DCNNs manage to do that—and an answer is a fine-grained one. Instead of dropping labels and all-or-nothing standards, the process of answering begins with the following transparent argument: DCNNs mimic the acquisition of general category representations through domain-general mechanisms, thereby vindicating empiricism. Here, we do not have grand claims about *tabula rasa* empiricism nor the endorsement of such empiricism. Rather, the discussion is about domain-general *versus* domain-specific mechanisms underlying higher cognitive processes. This means that Buckner does not deny that DL models contain priors *and* inductive biases (or any other serious philosopher who wants to defend empiricism, by the way). As he acknowledges regarding the AlphaGo Zero, the model was indeed provided with rules, rather than learning them from experience as per *tabula rasa* empiricism. However, that should not concern a moderate empiricist as long as domain-specific strategy heuristics are not part of the inner armamentarium, and if one reads carefully a paper by Silver and colleagues, they were not (Buckner 2018: 5360, fn. 13). Let us see how this translates to the case of abstraction in DCNNs.

A brief historical reminder would be helpful here. The two pertinent issues stemming from Locke's and Hume's views on abstraction are the following: how the mind learns to ignore mutually inconsistent features of exemplars to form an abstract representation of category membership, and how it selects appropriate exemplars for the same undertaking? In other words, it is unclear how the bi-directional mental travel—from exemplars to abstract categories and the other way around—actually comes about. Buckner thinks that the solution to both issues lies in the notion of abstraction-as-transformational-invariance, i.e., abstraction understood as a subpersonal visual processing of systematic transformations of perceptual similarity space (2018: 5345). Moreover, DCNNs can provide us with the understanding of the dynamics of this bi-directional mental travel. Now, recall that DCNNs have exotic architectural features, namely *myriads* of layers, including *convolutional* and *pooling* layer where linear function of convolution and non-linear function of pooling is performed. When fed with perceptual input, vector representations of exemplars from the input form multidimensional vector space dubbed "perceptual similarity space" (Buckner 2018: 5347). The distance within such a space should parallel the degree of perceived similarities between exemplars, while a specific "manifold" is a region which marks boundaries of a category that encompasses several exemplars. Finally, the output should amount to the correct classification of the transformed input within a perceptual similarity space. The first transformation of the perceptual input happens through applying convolution, which results in the detection of relevant features of exemplars. The second transformation happens through applying pooling (most notably, max pooling), thereby leaving out irrelevant features of exemplars. Hierarchical abstraction reduces the complexity of the perceptual similarity space made of many exemplars by iterative transformation of vector representations of such space into much simpler format. This represents the solution to the Lockean problem of how mind comes to leave out some of the features while the salient ones "imprints" onto category membership (Buckner 2018: 5357).

However, to capture the prior probability that a feature is relevant or not for classifying exemplar as belonging to a particular category, the model needs to incorporate strong *domain-general priors,* such as max pooling, since a DCNN must learn to control for nuisance variables, such as size, position, or angular rotation (Buckner 2018: 5346). These variables can negatively affect the training and subsequent task performance of a DCNN by fixating different perspectives of a single exemplar as if they are multiple exemplars. This could then result in wrong classification since manifold, corresponding to distinct categories, end up tangled in a perceptual similarity space. DCNNs are successful at disentangling manifolds *precisely because of the priors* (Buckner 2018: 5359). This hardly sounds like radical empiricism. Moreover, it would be wrong to assume that transformational abstraction performed by a DCNN amounts to mere perceptual classification, thereby giving the impression that post-connectionist models are *déjà-vus*. Take AlphaGo Zero, for instance. The model implements a DCNN and its domain-general priors included, *inter alia*, game board configurations, which are not limited to either visual or tactile modalities, but, rather, were symbolic (Buckner 2018: 5360). The transformational abstraction performed by AlphaGo Zero had an unexpected creative ramification as well: Silver and colleagues (2017: 357-358) did, in fact, report that the model showed unorthodox and previously unrecorded strategy of playing Go.

In a nutshell, a solution to the Lockean problem was linked to *discriminative* capacities of DCNNs in tasks requiring the processing of unimodal and multimodal data. A solution to the Humean problem of the selection of right exemplars, or the other direction of mental travel, is a bit more challenging since it hinges on the process of "undoing" convolution to reverse feature values to their initial state in the perceptual input. In a word, *generative* capacities of DCNNs are to be evaluated here. Again, domain-general priors can be of great help, e.g., by bootstrapping the process of unpooling with prior probability estimates of the likeliest manifestation of features during the process of transformational abstraction. As Buckner puts it:

> "These discriminative/generative hybrids demonstrate that the opposition between the Lockean and Humean approaches to abstraction is unnecessary; mathematically, they are two sides of the same coin, and both are required for a full empiricist account of reasoning [...]" (2018: 5362).

To see how generative capacities of DCNNs can be demonstrated, consider DeepArt AI tool, released in 2015, which allows for separating artistic style from the other constitutive elements of famous paintings from art history in order to "repaint" photographs. For each convolutional layer that recorded abstract information about the content of paintings, a correlational feature map was designed to parallelly record abstract information at a higher level regarding the manner in which the content was depicted (Buckner 2018: 5362-5363). In this way, even when one feeds new input, such as photographs, a DeepArt model is able to depict objects from the photograph in different artistic styles, e.g., as belonging to Impressionism or Cubism.

So far, according to steps (1)-(2) in Childers and colleagues' argumentation, Buckner has flagrantly betrayed empiricism and acknowledged domain-general priors even in models trained in unsupervised fashion. He also claims, in line with step (3) in Childers and

colleagues' argumentation, that transformational abstraction is differently performed in mammalian visual neocortex. To put it bluntly, DCNNs rely on backpropagation algorithm, manually labeled data, and static exemplars (at least in models that Buckner considers in his 2018 paper). These are all biologically implausible aspects of visual processing when compared to human and animal perceptual systems. Thus, Buckner openly admits limitations of post-connectionist models that nativists tend to reify in order to discard them completely. The difference is, however, his (justified) belief in the possibility of gradual technological progress of which DCNNs and successful simulation of transformational abstraction is just a glimpse. One of the interesting solutions he proposes for making DCNNs more biologically plausible is a multi-dimensional source of reward signals through which ANN performance would be closely linked to the environment (Buckner 2018: 5364). Midjourney and DALL-E, generative AI tools for boosting creativity based on DL and multimodal ANNs that have been released in 2022, can handle both textual prompts as input specifying the content and required images as outputs. In other words, these tools function similarly to DeepArt, with the sole exception of being multimodal rather than unimodal, and all it took was to be patient for some seven years to see such an important scientific and industrial technological innovation. I will return to these tools in the following Sect. when discussing the prospects of simulating semantic competence in post-connectionist models through linking DL models to embodied approaches to cognition in particular.

From a philosophical point of view, however, what remains unsettled is whether biological implausibility undermines explanatory ambitions of empiricists, as well as how many domain-general priors one can acknowledge to remain in the empiricist coterie. Since the latter issue is relevant for developing moderate empiricism as a position, I will leave it for the next Subsect. As for the former issue, the moment has come to integrate all points made earlier in the Ch.

At the end of his paper, Buckner (2018: 5367) proposes to regard DCNNs not as full-scale models of mammalian visual cortex that could serve for virtual brain analytics, but rather as idealized mechanism sketches that explore computational constraints of visual processing based on the assumed structural isomorphism between ANNs and biological NNs. In my view, this is a rather confound view on what DCNNs can offer us, but I see it as Buckner's building of defensive position in the wake of the impending attack from both nativists and philosophers of science when it comes to the explanatory power of highly idealized and opaque models. Nativists, as per their virtually unattainable standards we have encountered in Marcus (2018b) and Pearl (2021), value only clear cases of models with maximal explanatory power that yield all-or-nothing descriptions of cognitive phenomena. Hence, either we can see right through the model's inner mechanisms and outer mechanisms of a particular phenomenon, or we know nothing since we do not grasp casual underpinnings of such models in terms of innate machinery. This aligns well with some philosophers of science thinking that only deductive-nomological explanations are adequate explanations that deepen our casual understanding of phenomena. Buckner is, therefore, obliged to bite the bullet and admit that such explanations—going hand-in-hand with the maximal explanatory power—are not compatible with DCNNs. He makes an

excellent point about computational constraints stemming from the common structural features of artificial and biological visual processing. As I have been arguing in Sect. **3.1.**, DCNNs are currently prototypes of post-connectionist models offering how-plausibly explanations because of their (original and ever increasing) biological plausibility, or in Buckner's terms "structural isomorphism". Thus, explanatory power comes in grades rather than in all-or-nothing fashion: not every explanation must be a how-actually explanation if it is to be explanatory at all.

Where I depart from Buckner is the exact stage of mechanism discovery within models implementing DCNNs. Having a sketch of a mechanism presumes that we are in the dark regarding all constituents and their interactions within a functioning mechanism, and we have only a glimpse of possible mechanism organization given some constraints and idealized character of a model. This overly humble view was applicable to precursors of DCNNs, such as Fukushima's *Neocognitron*, through which he explored to what extent cat visual cortex is computationally adequate for modeling performance on some computer vision tasks. Contemporary DCNNs are much more akin to mechanism schemes since we do know a thing or two about couple of constituents and interactions (think back on functions of convolution and pooling), although we are still unsure about the degree of biological plausibility that can be incorporated without losing computational efficiency (i.e., giving up on the biologically implausible albeit efficient backpropagation). Ideally, the more details pertaining to mammalian vision we add, the greater are chances of obtaining full-blooded how-actually explanations that would (once and for all?) silence the ones obsessing over such explanations. The progress from different stages of mechanism discovery as well as the gradual increase in explanatory power that accompanies it clearly show the explanatory dynamics that is characteristic of connectionism from its inception. Models change with respect to their complexity and the faithfulness of simulations of human cognitive processing. Assuming and then "proving" their paradoxical stagnation and unchanging nature in terms of remaining limited in a same way as they always were despite obvious differences among contemporary and early computational architectures is not much different from Cesare Cremonini's refusal to look through Galileo's telescope to see mountains on the Moon which would empirically disprove his *pro*-Aristotelian geocentric convictions. *E pur si muove*!

To wrap it up, AI researchers inclined to empiricism may tend to flaunt their models given their industrial or scientific success in one domain, whereas the rivals inclined to rationalism may tend to oversimplify the debate altogether by taking flaunting at its face value and disregarding the other tacit merits of both models and methodology that they are criticizing. What happens next is that the other side does the same: empiricists take rationalist flaunting of scientific rigor at face value and strive to show that the success of their models is valid for multiple domains by enlarging the list of parameters, hyperparameters and other exotic architectural features. Both parties seem to value only how-actually explanations and regard DL model as being explanatory static rather than dynamic. It is as if each model is a form of crucial experiment that should tell us which computational modelling methodology is the only game in town, and ultimately, who wins

the Brain Wars.[59] Unfortunately for all parties, however, research program such as connectionism cannot be evaluated via crucial experiments—as is the case with scientific disciplines in which both *explanans* and *explanandum* are subject to the ongoing research due to our lack of understanding either of the two parts needed for an explanation of phenomenon. Inasmuch we are unsure about the type of computational architecture that fits, we are uncertain about a plethora of relevant details regarding the cognitive architecture which is to be simulated by a computational good fit. To put it even more strongly, no crucial experiment is possible (for the time being at least) precisely because of the explanatory flux in which post-connectionist models thrive.

How Many Priors are Empiricists Allowed to Accept? A Moder ate Neo-Empiricist Dogma

Specific hypothesis II stated that labeling a philosopher or a scientist as a rationalist or empiricist in the 20th and 21st century has a different connotation than it had in the history of philosophy because it is dependent on the additional theoretical commitments that they implicitly or explicitly assume about the nature of human cognitive processes. These theoretical commitments were tightly intertwined with engineering breakthroughs in the 20th century, whereas currently they are linked to the buffet of methodological and architectural perks of myriads of different models. In the Subsect. above, my aim was to show radical empiricism and radical nativism as different sides of the same coin. When put forward by scientists, e.g., AI researchers or cognitive scientists, these labels become mere instruments for occupying spotlights rather than thought-through positions imbued in philosophical legacy of British empiricists or Continental rationalists. Moreover, they become "mock positions" which hardly anyone can defend on rational grounds. Defending moderate empiricism and moderate nativism is a more feasible endeavor. When put forward by philosophers, they are indeed more nuanced, albeit often veiled in historical misinterpretations. Additionally, philosophers tend to skim over methodological details of DL models which can sometimes result in a confounding view of their explanatory power or the extent of their philosophical significance. The 21st century is the century of details when it comes to understanding both computational and cognitive architectures.

Here, I will be dealing with the main issue of moderate empiricism, namely, the amount and the kind of priors and inductive biases that can be allowed in a DL model and still discern moderate empiricism from moderate nativism.[60] Priors and inductive biases guide the process of learning of an ANN. When Yann LeCun, a radical empiricist himself, ended on ropes during a public debate with Gary Marcus in 2017, he admitted: "I would be happy at end of my career if we have a machine as smart as a cat [...] Or a rat." (cited in Hsu 2017). For his part, Marcus remarked that "[a] little innate structure might help you get a long way towards that", to which Cun replied with "[a] minimal amount of it, yes" (cited

[59] Or Linguistic Wars, or AI Wars, or virtually any war waging between the two opposing accounts of the nature of human mind across disciplines involved in the cognitive science hexagon.

[60] The question from the title ("*How Many Priors...*") was thrown at me after my talk at the *EECP Annual Workshop for Early Career Scholars* in 2020. I am grateful to Juraj Hvorecky for making me think about his question and the adequate answer to it.

in Hsu 2017). To a reader unfamiliar with the earlier work and media appearances of this famous pair of archnemeses, this could seem like virtually the same position. When Buckner advocated moderate empiricism in his 2018 paper, he sounded similarly almost repentant for adhering for such a position given that ambitions of hardcore empiricist like Silver and colleagues were unfulfilled.

The first distinction that comes to mind when trying to draw a line between moderate empiricism and moderate nativism is the number of domain-general and domain-specific mechanisms postulated within a model, and by extension, within the human mind. According to Margolis & Laurence (2013, 2015), the two coteries differ with respect to styles of learning-based explanations in the sense that empiricist insists on a small number of mechanisms that are generally and repeatedly engaged to process different inputs, whereas nativists think that a plethora of pre-wired specialized modules for different inputs guides the development of our cognitive capabilities. Gabe Dupre enhances this picture with the motivation behind relying on the numbers for the purpose of delineating:

> "On the empiricist view, different domains of knowledge (language, mathematics, social reasoning, etc.) are differentiated only by the information from which they have been generated: the mechanism for their acquisition is the same, general-purpose, learning system. [...] For the nativist/rationalist, on the other hand, if each of these domains is a product of distinct acquisition mechanisms, we will likely have to appeal to these distinct mechanisms [...]" (2021a: 1018).

The difference is, therefore, grounded in the number of domain-general and domain-specific mechanisms (in terms of priors and inductive biases) that one is allowed to incorporate into a computational architecture if one wants to make an inference from the functioning of the computational architecture to inner workings of a cognitive architecture. Besides mechanisms, the number of inputs also plays a significant role. Domain-specific and innate mechanisms will be triggered by a small number of inputs or stimuli that are precise, whereas domain-general mechanisms will be triggered by a wide range of inputs.

The domain-general/domain-specific dichotomy is often implicit in moderate nativists' explicit list of computational primitives. During his debate with LeCun, Gary Marcus proposed and later published such a list. He insists on the following (2018b: 12):

(a) Representations of objects,
(b) Structured, algebraic representations,
(c) Operations over variables,
(d) A type-token distinction,
(e) A capacity to represent sets, locations, paths, trajectories, obstacles, individuals,
(f) A manner of representing the affordances of objects,
(g) Spatiotemporal contiguity,
(h) Causality,
(i) Translational invariance,
(j) A capacity for cost-benefit analyses,
(k) Representation of time,
(l) Intentionality.

This sure is an extensive and impressive list. However, Marcus does not distinguish domain-specific mechanisms or capacities from domain-general ones, nor priors from inductive biases, but rather gives the impression that (*a*)-(*l*) must simply be innate or else DL models are good for nothing. Among them, there are indeed both kinds of mechanisms and capacities. For instance, (*a*), (*e*), (*f*), (*g*), (*h*), and (*k*) are linked to core cognition, or, more precisely, intuitive physics developed from infancy, and can be considered domain-general, although the format of (*a*) is for Marcus probably symbolic rather than distributed as in connectionism. On the other hand, (*b*) and (*c*) are a strong theoretical commitment that cannot be incorporated into DL models if the banners of moderate empiricism are to be defended given that such innate mechanisms are conceived as domain specific. Moreover, (*b*) and (*c*) cannot be incorporated into any post-connectionist model if connectionism strives to be accepted as an autonomous research program pertaining to human cognition. Thus, Marcus demands the impossible. Finally, (*j*) and (*l*) are capacities belonging to intuitive psychology, and (*l*) especially comes with a heavy philosophical baggage in which most AI researchers are simply not interested.[61]

Arguably, Marcus' is cleaving closer to moderate nativism instead of radical nativism, despite having an attitude, at least judging by the similarity of his list to desiderata that Melanie Mitchell, a moderate nativist in the field of AI engineering, proposes in her recent book. Mitchell acknowledges intuitive physics as cornerstone of cognitive development—even our concepts have physical basis, therefore, DL models are "blind" and worse off than any toddler we know without manually programmed inductive biases that would mimic our core cognition (see 2019: Ch. 4). In other words, conceptual knowledge is out of reach for post-connectionist models without augmenting the innate

[61] Moreover, some philosophers, most notably John Searle, would deny that it is *in principle* possible for a computer to have intentionality. He famously illustrated his argument with the *Chinese Room* thought experiment (1980, 1984) that purported to show how syntax (symbol manipulation) is not sufficient for semantics (genuine understanding of sentence and word meaning). Although the argument primarily targeted GOFAI and CTM, it is relevant for *any* artificial system since Searl essentially claims that computation is defined formally whereas our mental states have semantic content (viz., "aboutness") in virtue of causal powers of our brain that ensure connection between content and environment. I will not be opening a discussion here about Searle' *Chinese Room* given that there has been a tremendous number of publications addressing it—counting only the last decade, from 2010 to 2019, Google Scholar gives over 2000 different results (for an overview see Cole 2020). Needless to say, Searle would say the same thing for LLMs, i.e., that syntax is not sufficient for semantics, therefore LLMs cannot *in principle* ever understand meaning as humans do. My intention in Sect. **3.3.** is *not* to prove him wrong. My initial assumption is that NLP models can help us understand human linguistic competence by simulating it to some extent, not that NLP models *are* linguistically competent. In Subotić (2021b), I have argued that cognitive scientists inasmuch AI researchers need not be bothered by intentionality and the folk psychological conviction that mental states must be expressed by propositional attitudes, i.e., that they have semantic content in propositional format. Connectionist and post-connectionist models offer us a level of description that does not hinge on the intentionality because it is subsymbolic, so the intentional idiom can be eliminated without adhering to strong eliminative materialism *à la* husband-wife team Paul Churchland (1981) and Patricia Churchland (1989). At the same time, the intentional idiom can be legitimately used in other scientific disciplines in which the level of description does hinge on the intentionality, e.g., in history we are interested in reasons and motives of historical figures. In this Subsect. I use the distinction subpersonal/personal to account for virtually the same idea, although without going into details regarding the so-called *eliminativism of the limited domain* that I have introduced in 2021b.

armamentarium. Moreover, conceptual knowledge and innate intuitive physics allow for zero-shot learning, which means that toddlers do not need as much training data as DL models do. *Ipso facto*, the understanding of the content in, say, DCNN's or transformer's input layer is out of reach—all we have is a merciless, stochastic data cruncher that is better at predicting patterns than us, but ultimately fail to use such patterns in a meaningful way, i.e., to set goals or have any kind of preferences. Connectionism cannot get off the ground since it has no semantic engine.

The issue at hand is whether moderate empiricist can show that intuitive physics can be incorporated in a DL model, albeit with less *priors* than Marcus and Mitchell would presuppose. This would, in turn, show that moderate empiricism with respect to conceptual domain of knowledge can do with *some* priors, i.e., the indispensable minimum of domain-general mechanisms, and, ultimately, set the ground for developing further argumentation about some kind of semantic engine for connectionism. It is customary in the literature to lump together inductive biases and priors even though their functions within a model differ. Inductive biases relate to how ANNs learn, whereas priors amount to assumptions regarding the distribution of data, architectural features, and explicit or implicit knowledge about data interactions. Neither moderate empiricists nor rationalists deny that DL models have inductive biases since these include basic architectural features such as distributed representations, depth, convolution in the case of DCNNs, recurrence in the case of SRNs and RNNs, or self-supervised pre-training in the case of transformers. In the words of Lappin & Shieber, the linguists who discussed the relevance of ML for language acquisition as early as the first decade of the 21st century, "there is no such thing as no bias model" (2007: 2). However, what they dub "weak bias model" or "strong bias model" essentially means model with smaller or larger number of task-specific or task-general priors. Regardless of the degree of imprecision among scientists and philosophers, when assessing whether DL models show that intuitive physics, or any other knowledge domain, is aligned with empiricist or rationalist claims, we are interested in the number of priors and appreciate the transparency about the inductive biases. But one should bear in mind that not all priors are inductive biases, nor inductive biases are intrinsic to the model as manually specified prior is.

The issue of intuitive physics has, in fact, been tackled recently—some three years after Mitchell and Marcus voiced their position. Again, the moral is that the exploratory role of post-connectionist models should not have been underestimated. Piloto et al. (2022) have presented a DL model incorporating the mechanisms for intuitive physics concept acquisition underlying infant visual cognition. The team combined developmental psychology with engineering and through the violation-of-expectation paradigm probed a schema of a mechanism in a DL model. The violation-of-expectation paradigm is frequently used in developmental psychology as the working hypothesis is that possessing any physical concept involves forming a set of expectations regarding future events (Piloto et al. 2022: 1257). Thus, one can distinguish between cases when infants presented with visually similar arrays choose physically possible or physically impossible arrays, i.e., those that are either consistent or inconsistent with the specific concept. For instance, if infants, and/or adults, possess the concept of object *permanence*, they anticipate that objects will not

disappear when they are not in sight. Similarly, if they expect objects to avoid merging with one another, this indicates that they possess the concept of *solidity*. Moreover, the expectation that objects will follow continuous paths through time and space demonstrates the understanding of the concept of *continuity*. Piloto and colleagues have designed probe videos, each video demonstrating inconsistency with intuitive physics, which are paired with corresponding ones that establish a baseline that aligns with principles of physics. Additionally, they have also designed Physical Concepts dataset with 300,000 videos as training material for ANNs. The goal was essentially to find under which *generic constraints* their DL model *learns* to track objects so that it could be compared to human or infant performance in further iterations to see what mechanisms enable such capacity. As they report, such model is endowed with the perceptual module and the dynamics predictor: the former component (a CNN) converts perceptual input into a set of codes for objects and the latter component (a LSTM) predicts the future video frame based on the codes (Piloto et al. 2022: 1259). As long as the model remains object-centered, it can develop robust violation-of-expectation effects with a strikingly small amount of training data—*cca* 28h of visual experience. When tested on the unseen objects and events, without retraining or fine-tuning, the reported robust effects have not changed.

Regarding the number and kind of priors, Piloto and colleagues admit:

> "[I]t is important to acknowledge that our model implementation was granted access to two sources of privileged information: object (segmentation) masks and object indices allowing consistent placement of object embeddings within the model over timesteps (tracking). Recent research has introduced methods for object segmentation and tracking that would allow each of these to be extracted from pure video data without access to privileged information of any kind" (2022: 1263).

This quote illustrates nicely both the exploratory role of DL models and explanatory dynamics arising from that role. Researchers are evaluating the prospects of post-connectionist models having intuitive physics like any infant and toddler does through a quantitative method borrowed from developmental psychology and combined with DL, which makes them one of the pioneers. In this way, we can see how DL models can be used to empirically probe hypotheses pertaining to infant and adult cognitive processes. Besides, we can also savor the moment while witnessing how DL models can profit from the corpus of tools and experimental results of already established scientific disciplines, as well as use such results as independent evidence in favor of particular working hypotheses. Furthermore, by acknowledging that there are currently preprints in which modeling of intuitive physics can be even taken one step further with minimizing the priors, we can appreciate the explanatory flux of DL models that aligns well with the core tenets of empiricism and the optimism of the proponents of ANNs. Important thing to note here is that Piloto and colleagues explicitly and transparently state that their model is successful in virtue of having *two* priors (along with inductive biases such as convolution, distributed representations, and DL algorithm, *cf*. Goyal & Bengio 2020: 5). Strictly speaking, these priors do not amount to object representations as advocated by Marcus and Mitchell but do point out that faster learning comes with some kind of innate machinery, albeit radically different than the concepts of intuitive physics, which are—arguably—hardwired into our

cognitive apparatus. This radical difference allows for advocating a moderate position, namely that of moderate empiricism which is comfortable with a small number of domain-general priors and inductive biases that facilitate learning of intuitive physics but not with domain-specific conceptual representations.

Simply put, there are two distinctive characteristics of moderate empiricism:

(*A*) the endorsement of *domain-general mechanisms* (priors and inductive biases),

(*B*) *wide application* of such mechanisms.

The two related goals of moderate empiricism are the following:

(*A'*) to *decrease* the number of mechanisms as much as a model and task performance allow,

(*B'*) to *increase* the multimodality of ANNs to support the wider application.

Most likely, the successful simulation of human cognitive processes must rest on the combination of several complex ANNs like Piloto and colleagues interlace a CNN and a LSTM. Their successful simulation of intuitive physics, however, has no bearing on the semantic processing, since the model itself does not relate to such capacity. What can be inferred from it, however, is that the greater ecological validity and richness of perceptual data can aid in model's better acquaintance and manipulation of the content, which would, in turn, makes empiricists' engine going when it comes to the ambition that connectionism can encompass semantic cognition as well. In other words, the model has semantic potential, although the burden of proof is on empiricist modelers to show in what sense the potential can be realized within ANNs, and as I will be pointing out in the next Sect., there is no shortage of interesting post-connectionist models of semantic processing in this regard. For the time being, the main point is that conceptual knowledge in the domain of intuitive physics can be simulated in a post-connectionist model having a small number of priors of which neither is domain-specific and neither encodes a representation in a format that moderate rationalists would expect and require.

Cameron Buckner rightly remarks that

"[A]mong the relevant questions today is [...] not whether empiricists are allowed any innate structure or architecture whatsoever, but rather how much domain-specific knowledge can be extracted from incoming sensory experience using biased domain-general transformation methods, and how much [...] requires domain-specific programming whether by genes into brains, or by computer scientists into silicon" (2023: 20).

Thus, another important and relevant example of the role that priors and inputs play in the distilling of moderate empiricist and nativist positions is Chomsky's universal grammar. Contrary to the case of intuitive physics, where the number of domain-general priors was decisive for *domain-general knowledge*, the main concern now is whether domain-general priors can generate domain-specific knowledge. Also, conceptual knowledge is not in focus here, but rather *structural* knowledge. As described in Sect. **2.1.**, Chomsky (1957/2002, 1965, 1966) started with the idea that we have an internal set of a great number

of grammatical rules, or a set of discrete parameters, that help us master and understand language. During his career, he kept reducing the number of innate mechanisms that constitute our linguistic competence: first to a scintilla of principles and parameters (in 1980), then to *Merge* & *Move* operations (in 1995/2015), and, finally, to recursion (in 2002, 2015, 2016). Moreover, he has even admitted that the ability to have recursive thoughts and language applies across different domains, although being, at the same time, a fundamental language-specific constraint.[62] This is because there is a difference between the faculty of language in the broad and narrow sense. The former comprises sensory-motor and conceptual system (akin to semantic competence in Sect. **4.3.**), while the latter comprises recursion (akin to syntactic competence in Sects. **4.2.** and uniquely human and unique to language mechanism that generates an infinite set of hierarchically structured expressions yielding interpretations at the interface of sensory-motor and conceptual system (Hauser et al. 2002: 2-4). Thus, Chomsky started as a radical nativist about grammatical structures (with all the support from historical rationalists he could gather) inasmuch Fodor was radical nativist about concepts, and ended up as a moderate nativist whose commitments are distinguishable from moderate empiricism only to the keen eye. What has remained the same, are, however, the assumption that syntax is *cognitively autonomous*, regardless of whether its origin lies in the universal grammar or recursion, as well as the assumption that linguistic representations are *structure-sensitive*.

Even though the domain-specificity of language has been shrinking during Chomsky's career, this has not impeded proponents of symbolic cognitive science and GOFAI, like Gary Marcus, from insisting on the hierarchical nature of language. Thus, Berent & Marcus claim that without fundamental commitments of TGG

> "[t]here is no hierarchical organization of sentences, morphemes, or syllables; such formal constituents play no causal role in mental processes [in connectionist models]. Instead, learners only extract the statistical structure of the lexicon. Productivity, then, is limited to lexical analogies; no linguistic analogies can extend across the board" (2019: e82).

Additionally, the open-ended nature of language suggests that humans ought to map between a potentially infinite range of input sentences and meanings to communicate, many of which they probably have never encountered, and this could not be possible without domain-specific constraints, such as recursion. For Marcus (2018a), this clearly shows that we generalize beyond "training space" quite routinely as opposed to DL models. Note that "productivity" and "generalization" are tightly intertwined with the systematicity of ANNs, only that the former term is common for linguistic, while the latter

---

[62] In general, linguists distinguish between two types of recursive structures, namely tail and complex. Take, for instance, sentences (S_1) The flea bit the horse that chased the dog that ran away, and (S_2) The dog that the horse that the flea bit chased ran away. Both (S_1) and (S_2) have roughly the same semantic content. Nevertheless, in the former sentence, the two levels of tail recursive structure are easily comprehended, while in the latter, the two levels of center-embeddings are difficult to process. Chomsky thinks that both types of recursion are built-in mechanisms of grammar even though complex recursion is almost never produced (see, e.g., Karlsson (2007) who reports the results of corpus analysis that show the absence of center-embeddings in spoken Swedish, Danish, English, German, French, Latin, and Finnish). Of course, Chomsky (and any other TGG proponent) has a relatively simple solution to this, namely the competence vs. performance distinction. As I will be arguing later, this is *not* a flawless solution.

for ML community. Rawski & Heinz (2019) back up this with the computational learning theory, specifically with the No-Free Lunch Theorem stating that no system could generalize beyond the training dataset unless that system has some prior knowledge that constrains the structured "hypothesis space". Simply put—you got to pay in order to eat. Rawski & Heinz accuse the proponents of DL models that their "strategy is simply to go for broke and learn everything", which means that, according to them, it seems that any pattern is learnable with enough data, including linguistic patterns. However, as Rawski & Heinz claim, this would be highly linguistically incorrect since linguistic generalizations attested in many world languages (viz., linguistic universals) are not arbitrary, but restricted by the innate domain-specific mechanisms as TGG predicts.

Things, therefore, look quite bleak for moderate empiricism in the case of domain-specific knowledge. On the one hand, the generalization beyond training data seems to be out of reach for DL models without strong domain-specific priors. The underlying linguistic rules and structure-sensitive representations enable speakers to combine words and linguistic expressions in systematic ways to create new expressions that are intelligible within the language system. Without such systematicity of language, productivity would be limited, as there would be no consistent rules to guide the creation and interpretation of novel, previously unseen expressions. Thus, DL models are in this regard faced with being psychologically *and* linguistically implausible—at least judging by the TGG framework. The linguistic universals as postulated by TGG and one of its core assumptions that productivity is an essential feature of both thought and language attest that NLP in connectionist models is relevant for our linguistic competence as much as novels about utopian societies are for effective institution building. An even stronger point can be advanced: DL models could not have simulated linguistic capacity in the first place since they never encoded structure-sensitive representations. Without such representations, there is no way to account for systematicity and productivity.

Luckily for moderate empiricism, Cameron Buckner inaugurates the new empiricist dogma in his forthcoming book *Deeply Rational Machines*, which also aligns with and reinforces the two main characteristics and goals of moderate empiricism that I set out at the beginning of the Subsect. The neo-empiricist dogma—or in Buckner's preferred spelling *DoGMA*—presumes that both higher and lower cognitive processes can be modeled and simulated by a domain (*Do*)-general (*G*) modular (*M*) architecture (*A*).[63] The negative side

[63] I use the standard spelling to stress the continuity between Quine's famous take on the two dogmas of empiricism (Sect. **1.3.**) and the new dogma. The implications of this continuity will be mentioned in **Conclusion**. Also, Buckner frames the book in terms of reviving the notion of psychological faculties and juxtaposes his DoGMA to the eliminativism about faculties proposed by neuroscientists inclined to radical empiricism such as, e.g., Pessoa, Medina & Dessfilis (2022), or, traditionally, the Churchlands (1989, 2007, 2013). I remain neutral about this issue and discuss mechanisms rather than faculties since I am specifically interested in mechanisms governing linguistic competence rather than in the taxonomy. Furthermore, my aim in the thesis is far less ambitions than in Buckner who explores all faculties that he deems relevant for artificial general intelligence that can be contrasted to capacities and mechanisms (*a*)-(*k*) in Marcus (2018b), and spells constraints and requirements for the computational architecture that could subtend such faculties (see 2023: 39-41 for a brief overview). The level of mechanisms allows me to stick to the subpersonal level of description

of the dogma amounts to prohibiting innate concepts or domain-specific heuristics, whereas the positive side sets forth the project of deriving abstract knowledge of several domains from architectures comprising multiple different ANN modules characterized by diverse learning algorithms (Buckner 2023: 28). Modules in Buckner's novel dogma are nowhere near similar to Fodor's (recall Sect. 2.2.) since they are not domain-specific and informationally encapsulated, although the information-sharing in them is more internal than external. Additionally, Buckner admits that they may but do not have to correspond to neuroanatomical brain regions. The dogma is essentially an empirical hypothesis about the adequate computational—and by extension about cognitive—architecture:

> "[...] if there are abstract correspondences between the structure of a neural network and the structure of human brain or mind, then the study of the computational model might reveal how humans actually do it, and more specifically that we actually do it without innate domain-specific knowledge" (Buckner 2023: 31).

Judging by what Buckner says, the dogma is as explicit and resolute about the tenets of post-connectionism as The PDP Bible was about "shallow" connectionism. My primary aim is to show how linguistic competence can be encompassed by the new empiricist dogma in line with characteristics (*A*) & (*B*) and related goals (*A'*) & (*B'*).

First, I plan to show that there is enough evidence in favor of the moderate empiricist expectation that language emerges from domain-general cognitive constraints and enough evidence that casts doubt on the specific assumptions of TGG such as linguistic universals being cornerstone of linguistic generalization. I focus on debunking the recursion as the proposed domain-specific mechanism. Unlike intuitive physics which constitutes our core cognition, the origin of linguistic competence is a matter of a heated debate in linguistics, that often goes under the radar of cognitive scientists and AI researchers. The distorted image, lurking in papers by nativists like Marcus, Rawski & Heinz, or Chomsky himself, shows the theoretical framework of TGG as the only one that can be reliably taken as reflecting our true and unique capability for language processing, thereby making any computational model that does not encode whatever features the framework presupposes obsolete in accounting for our cognitive architecture. Furthermore, a straight line can be drawn between early and post-connectionism precisely because of the reluctance of TGG aficionados to admit that human linguistic capacity can *ever* be simulated faithfully enough with ANNs (recall Sect. 2.3.) In other words, I first settle the overdue debt for lunch. In Sect. 4.3., I will take bull by the horns and deal with the structure-sensitivity of linguistic representations and the systematicity conceived as the essential features of both thought and language.

Christiansen & Chater (2015) compile evidence from studies about language evolution, cognitive neuroscience, and connectionist modeling to argue that our capability to process recursive structure is not related to recursion *qua* faculty of language in a narrow sense but originates from domain-general mechanism governing sequence learning. In

---

where I can eschew the issue of eliminativism vs. realism about faculties in a similar manner as I eschew it regarding intentionality (see fn. 61). Truth be told, I do indeed lean towards eliminativism (as in Subotić 2021b).

doing so, they indirectly defend UBT as a rival framework to TGG. First, Christiansen & Chater (2015: 25) point out that recursion, despite being considered a linguistic universal within TGG, is not a feature of every single language from world's current 6-8000 spoken languages[64], and recursive structures are far from uniform in them since there are significant individual differences related to processing difficulty, which, in turn, affect the overall distributional structure of these languages. The individual differences include but are not limited to, say, level of education that correlates with faster and somewhat easier processing of complex recursive structures (Dabrowska 1997), and general experience with specific language patterns (Christiansen & Chater 2016: Ch. 7). This is the core tenet of the usage-based theories (UBT), namely that language consists of acquired constructions that are essentially generalized patterns scaffolded by our joint communicative practices, cognitive capabilities, and environmental cues (Tomasello 2003).[65] In this sense, UBT denies that there are any linguistic universals as postulated by TGG, rather universals are probabilistic trends in language usage, not rigid structures originating in the domain-specific prior such as recursion (Christiansen & Chater 2015: 35). Hence, the capability to handle *recursive structure* (i.e., repeated usage of the same construction)—if there is any in a particular language—does not hinge on *recursive domain-specific mechanisms*. These are two completely different things. Instead, this capability should be seen as an acquired skill, learned through the qualitatively and quantitatively rich or poor experience with instances of recursive constructions. In other words, such capability has nothing to do with "competence grammar", but

> "[p]erformance limitations emerge naturally through interactions between linguistic experience and cognitive constraints on learning and processing, so that recursive abilities degrade in line with human performance across languages and individuals" (Christiansen & Chater 2016: 203).[66]

---

[64] Probably the most controversial dispute in this regard is the status of Pirahã language. Daniel Everett (2005) was doing fieldwork in the Brazilian state of Amazonas and discovered that members of the Pirahã tribe do not use recursion since they do not use subordinate clause at all (i.e., they do not combine or expand sentences), nor expressions for quantity, nouns designating abstract concepts, pronouns, etc. The reason for this peculiar mode of communication lies in the custom of the tribe that one is allowed to convey one's direct and immediate experience. This can happen not only through language but also dance, gestures, music, or chirping. Everett's conclusion was that recursion cannot be seen as a prerequisite for language capability since members of Pirahã do think recursively but do not have recursive language. Moreover, even when Pirahãs learn Portuguese, a language that has recursive structures, they usually speak it non-recursively unless they were brought up in the urban Brazilian areas and cut off from traditional customs from the early age (Sakel 2012). Everett (2017) extensively wrote about the "intellectual prosecution" he faced in linguistic circles for trying to conclusively refute Chomsky's TGG, whereas TGG aficionados either insist that Pirahã case is irrelevant or that Everett is a charlatan. Nonetheless, Evans & Levinson (2009) also listed several indigenous languages without recursive structures, further casting doubt on the alleged universality of recursion.

[65] I have much more to say about UBT in what follows. Teaser: I will argue that inasmuch TGG was matching theoretical framework for providing *Bauplan* of linguistic competence within the symbolic cognitive science, UBT along with other strands of cognitive linguistics plays the same role within connectionism.

[66] It is worth noting here that Chomsky's appeal to recursion was met with raised eyebrows from those that hurried to challenge the assumption of human exceptionalism. Corballis (2014) aimed to show that recursive thinking is also present in great apes, and ipso facto, that recursive thought can exist without recursive

What sorts of cognitive constraints on learning and processing do the authors have in mind? The constraints are spelled out by a domain-general, innate mechanism of sequence learning and processing, on which our capability to handle recursive structure piggybacked through our evolutionary history (Christiansen & Chater 2015: 27).

To support their argument that faculty of language in a narrow sense is—strictly speaking—empty, Christiansen & Chater draw on their earlier work in connectionist modeling. The SRN-based connectionist model of Reali & Christiansen (2009) was designed to *explore* how recursion could recruit complex sequence learning. First, a "biological evolution" was simulated through training around 500 generations of SRNs on simple sequence learning task by producing new generations from the "genome" (i.e., initial weights) of the most successful SRN in a given generation. Only then was the model filled in with a flexible grammar skeleton without any constraint regarding the exact word order and containing instances of recursive constructions. The processing of context-free grammar skeleton was constrained by values obtained on sequence learning task. This was seen as a "cultural evolution" and the result was that after more than 100 generations of SRNs, a consistent word order began to emerge. The overall conclusion was that recursion may originate during the cultural evolution in the absence of any domain-specific mechanism, but thanks to the domain-general mechanism such as sequence learning. The important thing here for proving my point that moderate empiricism regarding connectionist computational and cognitive architectures can accommodate language processing with domain-general mechanisms rather than domain-specific ones is that this conclusion was derived precisely from a connectionist model that had only a few priors for handling sequential learning as per Reali & Christiansen's hypothesis.

However, one could say that this model boils down to the mere how-possibly explanation that has no grounding in data pertaining to actual human performance on complex recursive structures. Because of this, the model may be taken to be a representative of the highly idealized performance inasmuch as TGG offers us a rosy, over-intellectualized image of competence that has no grounding in experimental and anthropological data (fn. 62 & 64). Consider, however, the following model. Christiansen & MacDonald (2009) modeled existing human data on the processing of central embeddings and cross-dependency structures.[67] They then compared it to an SRN specialized for sequence learning through Grammatical Prediction Error scores that reflect SRNs ability to predict the next word in sequence given the previously seen context and found out that both an ANN and humans exhibit similar qualitative patterns for processing difficulties of the two syntactic phenomena. Their conclusion was that the successful simulation was facilitated by a combination of intrinsic computational constraints such as manually specified priors

language. In a similar vein, my aim is to do show that the crucial thing about connectionism is that it rests on decoupling the thought from language. Both Corballis' and my points have an implication for the philosophy of mind that I tackle in **Conclusion** regarding the minimal conditions for ascribing personhood to nonhuman agents.

[67] A cross-dependency structures is a syntactic phenomenon, typical for Swiss-German and Dutch, where the processing of one element depends on multiple other non-adjacent constitutive elements in a sentence. This happens due to syntactic movement or displacement of constitutive elements within a sentence.

for sequential learning and distributional properties of the input to which both the SRN and humans were exposed since the training dataset included natural language grammar rather than context-free artificial grammar. With this model in mind, we can see the incremental progress regarding the explanatory prospects of connectionism: from a sketchy model that basically extended Elman's initial model to cover multiple SRNs to chart the space of possible constraints that could constitute a cognitive mechanism underlying particular cognitive process, connectionist modelers moved to a mechanism scheme calibrated with relevant details from psycholinguistics to show that "brain is not shaped for language, [rather] the language is shaped by the brain" (Christiansen & Chater 2015: 34).

In my view, further steps of augmenting the explanatory power should include increasing the psychological plausibility to come closer to how-actually explanations from the starting point of how-plausibly explanations. This can be done along the lines of Piloto et al. (2022) who rely on developmental psychology. For instance, Wang et al. (2023) train a DL model by capturing a subset of linguistic and visual stimuli encountered by a single child during the developmental process thanks to the longitudinal developmental dataset of 6 to 25-months old children that became available in 2021. As they put it, their work "provides a unique window into the learnability of linguistic structure based on one child's input—without additional data and labels, using a distributional learning strategy" (2023: 2). Wang et al. (2023) use a multimodal model that combines LSTM with CNN and train both in an unsupervised manner. Their results show that such a model learns to differentiate syntactic categories including nouns, adjectives, adverbs, transitive, and intransitive verbs, as well as subcategories of nouns such as animals, body parts, and clothing, thereby suggesting the roots of semantic competence and syntax-semantics interface at this particular developmental axis. Visual representations bring incremental progress in predicting the next words within a context, especially for verbs and nouns. The authors did not probe model's capability for generating recursive structures in particular, however, they did show a plausible way of marrying DL and developmental psychology for the purpose of advancing moderate empiricist take on syntactic and semantic processing by exploring the developmental constraints on syntax-semantics interface. The emerging syntactic categories within the multimodal combo of different ANNs that did not receive any fine-tuning or manual labeling of developmental data indirectly suggest that the ability for generating such categories may need not be domain-specific and in line with nativism.

A step towards how-actually explanation based on a model which would be trained on similar developmental data would be to combine the approach of Wang et al. (2023) in terms of multimodal ANNs coupled with detailed independent evidence from psycholinguistics. Thus, Poletiek et al. (2018) examined under what conditions complex recursive constructions, such as center-embeddings, can be learned and reported that length reduction fulfills a secondary function in comparison to complexity reduction, especially in child-directed speech, which essentially means that starting simple is more important than starting short. As per their study, starting small allows the learner to focus on simple patterns that make basic structural elements salient, which, in due time, facilitates mastering of more demanding patterns and, *ipso facto*, better performance. Wang et al. (2023: 12) admit that their training set is scant in comparison to children's linguistic experience:

the model runs on 225 thousand of tokens, while children receive *cca* 3-20 million of words per year during first two years of their lives. Thus, exposing gradually a multimodal model to more and more computationally complex sentences and juxtaposing its performance to a relevant and diverse amount of developmental data[68] along with relying on the goals (*A'*) and (*B'*) would represent a proposition of a genuine how-actually explanation of recursive syntactic processing that goes against core assumptions of TGG.

To sum up, after presenting the domain of intuitive physics, the domain of language acquisition served to show that the key difference between moderate empiricism and moderate rationalism boils down to the kind of encoded knowledge within DL models. A recent model proposed by Sartran et al. (2022) is based on the implementation of syntactic recursive structure to constrain several different computational architectures for NLP. It turns out that the stronger prior is, the almost perfect performance of a computational architecture emerges. This seems like an ideal model that would prompt Gary Marcus to say "Told you so"—just check again the items (*a*)-(*d*) from his list of computational primitives that I presented at the beginning of this Subsect. The model mimics the proposition that Marcus gave to many connectionists over the years: explicitly build in the hierarchical structure of linguistic expressions if you want to come any closer to the cognitive architecture that traditional symbolic cognitive science advocates for decades because it is the only legitimate hypothesis about our linguistic competence. The strong domain-specific prior such as tree structures and recursion are big no if one wants to defend empiricist banners and the autonomy of connectionism as offering both cognitive and computational architecture. Interestingly enough, moderate empiricist does not strive after the almost perfect performance because she does not begin with the conviction that linguistic competence is almost perfect, neat, and rule governed. The difference in amount and kind of priors in methodological terms has ontological ramifications in terms of the nature of linguistic competence and the status of competence vs. performance distinction.

[68] How much data is enough data in this case, though? I will get back to this question in the next Sect. since current LLMs implemented in chatbots are all about scaling, i.e., getting larger and larger, whereas some linguists (e.g., Huebner et al. 2021) urge that small-scaled child-directed language are more informative for understanding the nature of language faculty. In other words, we witness the friction between commercial/engineering and purely academic/scientific aims once again. Moreover, as I will be mentioning in the next Subsect. and **Conclusion**, besides quantity, the *quality* of data has an important role for assessing all faces of LLMs (Bender et al. 2021).

4. A TRIUMPH OF THE UNDERDOG: THE NOVEL ACCOUNT OF LINGUISTIC COMPETENCE

> *Overall, my linguistic competence is a product of sophisticated algorithms and large-scale language training, enabling me to assist users with a wide range of inquiries and engage in language-based tasks with human-like fluency and versatility.*
>
> – ChatGPT (17th May 2023)

4.1. The Stochastic Nature of Linguistic Competence

The ambition of a moderate neo-empiricist dogma is to show the viability of connectionism qua theory of human cognition by respecting the following credo: account for higher cognitive processes, that were out of reach for shallow connectionism, by relying on *deep*, *domain-general*, *modular,* and *multimodal* computational architectures. The time has come to examine whether the dogma encompasses DL models for NLP and LLMs and in what sense these models can be illuminative for our linguistic competence. In other words, we have reached the tipping point of this dissertation, from this moment on, either everything starts to make sense, or the sense is lost once and for all. Drama aside, this is the Sect. where the main hypothesis is addressed. Recall, the main hypothesis is that if one can show that linguistic competence can be *examined*, *explained* and *simulated faithfully enough* via models of syntactic and semantic processing, which are not based on the application of encoded rules and symbolic representations, but, rather, on DL and huge amount of data, then it is more scientifically fruitful and philosophically convincing to endorse moderate empiricist account of linguistic competence as opposed to rationalist account. Let me break down the main hypothesis into key claims I am about to defend.

(1) LLMs and DL models for NLP are used both as auxiliary tools and computational models on their own in accordance with their exploratory role. In order to show that linguistic competence can be examined by them, there should be independent empirical evidence that these models are relevant for such scientific pursuits beside their industrial application.
(2) In line with (1), some of the models should provide us with at least how-plausible explanations of some aspects of our linguistic competence. The prospects of offering such explanations suggest that it makes sense comparing LLMs or DL models for NLP and human cognitive processing and that these models can be informative regarding the mechanisms underlying linguistic competence.
(3) In line with (2), LLMs, along with other DL models for NLP (most notably multimodal, LSTMs, and RNNs), should be assessed with respect to their powers of simulating syntactic and semantic processing to see what kind of architecture is the most adequate for advancing the neo-empiricist dogma and promising for the development of how-actually explanations.

I will briefly cover (1) and (2) so that I can proceed to (3). While addressing (3), the focus will be on the *systematicity challenge* and *representationalism* so that I can directly answer criticism coming from the proponents of TGG and symbolic cognitive science. I will argue, *inter alia*, that if a product or output of a cognitive process has a feature φ, it does not follow that the very process must be attributed with φ. If I manage to establish that DL models of NLP can meet the systematicity challenge (unlike shallow connectionist models) and that the commitment to representationalism does not come at a high cost, then I have made a compelling case for connectionist cognitive architecture for linguistic competence, although it remains to be seen whether LLMs reinforce this case or not. Finally, I chart what can reasonably be expected from post-connectionist models of NLP and what is currently beyond their prowess. However, since I also believe that these models are future-biased, I will put my finger on the direction of their further development when it comes to linguistic competence.

First things first, what reasons do we have to consider LLMs similar to our linguistic performance? Huebner et al. (2021), who present acquisition-friendly model BabyBERTa as a resource for developmental psychologists, rightly remark that English-speaking children by their sixth birthday acquire near adult-like grammatical knowledge by being previously exposed to no more than 10-50 million words, which is *cca* 600 times smaller corpus than RoBERTa, a LLM trained on 800 million words of BooksCorpus and the full text of the English Wikipedia (*cca* 2.5 billion words). *Prima facie*, given the amount of training data and parameters (e.g., RoBERTa has 125 million parameters), LLMs seem like Godzillas of AI engineering.[69] However, two recent studies in neuroscience show that these Godzillas are surprisingly and exceedingly good in predicting semantic comprehension from brain activity (Caucheteux, Gramfort & King 2022) and that they share core computational principles with humans when it comes to neural encoding of language (Goldstein et al. 2022).

Thus, Caucheteux, Gramfort, and King examine the relationship between representations of GPT-2 (released in 2019 and superseded by GPT-3) and fMRI of human subjects' brain activations based on dataset that includes processing and comprehension of seven different short stories. The researchers decompose the architecture of GPT-2 to investigate the function and influence of attention heads, layer depth, and detailed task performance resulting in specific brain and comprehension scores that are being compared to human scores. Their results suggest that the increase in attention span allows for better prediction of text comprehension as well as that the deeper layers processing more contextually complex data are good at predicting the activity of frontal-parietal regions of brain (Caucheteux, Gramfort & King 2022: 16327/4). On the other hand, shallow layers deploying short attention span are good at predicting the activity of low-level acoustic regions of brain, so the researchers remark that there is a hierarchy of neural representations that is reflected in architectural features of GPT-2. In their view, LLMs could be made more brain-like by tweaking attention heads to allow for both short and long span depending on

[69] *Nota bene*: Both BabyBERTa and RoBERTa are based on BERT, one of the first LLMs to be released (Devlin et al. 2019), which I have briefly mentioned in Subsect. **Computational Architectures for the 21st century**. More information about BERT awaits in the Sect **4.2.**

a level of representation, be it lexical, sublexical (i.e., phonetic), or supralexical (i.e., syntactic) since the initial plausibility between story comprehension and mapping meaning to brain responses by GPT-2 suggests that DL models capture something relevant about the underlying mechanisms of language processing, although we are still unsure what exactly.

Goldstein et al. (2022) root such initial plausibility in the newly discovered behavioral and neural evidence pertaining to fundamental computational principles. They compare performance of GPT-2 and human subjects on the next-word prediction task by distributing a 30-minutes long transcribed podcast as training dataset and calculating the predictability score. As they report, human predictability score and GPT-2 estimations of predictability are highly correlated, and with the increase in contextual window size (i.e., the length of longest distance of tokens that a model can use to generate next token), the correlation coefficient rises as well—only 9.2% words that human predicted correctly were not predicted by GPT-2 (Goldstein et al. 2022: 371-372). However, they were interested in whether there is neural compatibility between the model and human subjects that governs this task performance in the background, i.e., they wanted to find out whether humans also engage in spontaneous word predictions before word onset. The researchers gathered neural data from epilepsy patients engaged in free listening to see whether language encoding is influenced by contextual dependencies as it would be the case in LLMs. Not only that Goldstein and colleagues (2022: 377-378) confirm the hypothesis, but also check for neural underpinnings of post-onset word surprise. This is because LLMs provide a unified framework for coupling pre-onset word prediction and prediction-error signals, so that they can safely infer that observed correlations corroborate the relationship between GPT-2 *qua* cognitive model and human task performance. Interestingly enough, the researchers explicitly juxtapose GPT-2 to traditional symbolic models used in psycholinguistics, and insist on the lack of interpretable, manually specified rules within a model. However, they also rightly remark that parallelized transformers such as GPT-2 lack full biological *feasibility* given that these ANNs account for essentially serial computation in humans. In my terms, the researchers acknowledge that they have not offered how-actually explanations, nor that they can promise them for the time being.

Both studies seem to resonate with Smolensky & Legendre's (2006) conviction that the brain is connectionist: when it comes to neural underpinnings and behavioral output, i.e., implementational aspect, connectionism is well positioned, even in the case of language comprehension. Nonetheless, the mechanisms behind language comprehension, which are constitutive of our linguistic competence, can perhaps be symbolic rather than connectionist in spirit. Again, in Smolensky & Legendre's (2006) register, there has never been a promising alternative to symbolic "mind" computation, especially when it comes to higher cognitive processes. Throughout past thirty years, there have been many ecumenist propositions from linguists—the most recent one coming from Joe Pater (2019)–and all of them relied on the strategy of allowing TGG and symbolic cognitive science to form a unified front as a paradigm for linguistic competence, with ANNs being just exotic and exploratory implementational tools of such a paradigm. On the other hand, in the most recent manifesto, Contreras Kallens, Kristensen-McLahlan, and Christiansen ask the following: "[...] If the overwhelmingly grammatical language produced by LLMs can be

explained by statistical learning and generalization, what is the need for innate grammar?" (2023: 3). In their view, LLMs are a UBT-friendly answer to the Poverty of stimulus argument since they bring to the table systematic explorations of what can and cannot be learned from the statistical regularities alone. In other words, LLMs show that our grammatical knowledge, which was considered domain-specific and rule-driven in TGG, has domain-general origins and probabilistic nature. In Sect. **3.2.**, I have argued for a similar point in the context of providing a moderate empiricist view of the origins of higher cognitive processes, such as language, and proposed a general connectionist and UBT-backed up account of recursion. In this Sect., my primary goal is to show how such account encompasses not only the origins of language, but both components of linguistic competence, namely syntactic and semantic competence.

Contreras Kallens, Kristensen-McLahlan, and Christiansen are researchers working in psychology and cognitive science, so for them the probabilistic nature of linguistic knowledge is something they infer from their preferred methodology, namely connectionist computational modeling. Linguists, however, have their own theoretical baggage to deal with, and it takes much more than computational modeling to establish the same inference in linguistics. Any theory in linguistics, be it TGG or UBT, has to account for grammaticality and acceptability, i.e., the difference between linguistic expressions that are grammatically correct and that sound "okay" or "okayish" to native speakers. Acceptability is measured by asking native speakers to rate sentences, and the working assumption is that they are able to do so due to the grammatical competence[70] that is taken to be a theoretical entity in linguistics. Grammatical competence should somehow generate a set of linguistic expressions that are well-formed and through which it is possible to discard all other expressions as ungrammatical. Acceptability, on the other hand, is often influenced by performance and processing factors, i.e., external and contingent factors. In a nutshell, many linguists endorsing this particular understanding of grammaticality and acceptability, and working within TGG, would say that grammatical competence has to be categorical, whereas acceptability can be probabilistic. Alternative view would be to see grammatical competence as generating a probability distribution over a set of linguistic constructions, including both well-formed, ill-formed, and something in-between (Lau, Clark, & Lappin 2016: 3). Linguists inclined to the alternative view usually work within UBT and think that central range of data that should be explained by a linguistic theory comes from our everyday usage of linguistic expressions, while TGG aficionados rely on artificial and formal grammar examples to set exact conditions that constrain everyday usage.

Lau, Clark, and Lapin (2016) purport to show that linguistic competence has probabilistic nature and can encompass both the stochasticity of actual language use *and* predict gradience of acceptability, which then can be compared to the outputs of inherently probabilistic post-connectionist models. Instead of relying on professional syntacticians' ratings of sentences, the researchers relied on crowd sourcing and bidirectional machine

[70] Grammatical competence is something I am about to dub "syntactic competence" to distinguish it from "semantic competence". Both constitute linguistic competence, as I will be elaborating in the rest of the Sect.

translation of 600 sentences in English to Norwegian, Spanish, Chinese, and Japanese in order to yield linguistic expressions of varying level of acceptability. For instance, in one of the probes, human subjects were to choose between four options, viz., 'extremely natural', 'somewhat natural', extremely unnatural', 'somewhat unnatural' and thus provide us with their acceptability judgments (Lau, Clark & Lapin 2016: 9). Regardless of the type of the probe, the results were promising since they suggested that acceptability judgments are intrinsically gradient rather than binary. However, this still does not prove that grammaticality is not binary, and that competence is not categorical. Lau and colleagues turn to unsupervised NLP models: "if [such model] can reliably predict human acceptability judgments, then it provides a benchmark of what humans could, in principle, achieve with the same learning algorithm" (2016: 17). In other words, if probability distributions of NLP models accurately predict acceptability judgments, and, *ipso facto*, explain their intrinsic gradience, then positing underlying categorical competence is redundant because human linguistic knowledge seems to be probabilistic inasmuch it is captured by a probabilistic model. In addition to the dataset used for the earlier behavioral study, Lau and colleagues include datasets based on English, Spanish, German, and Russian Wikipedia. The mean ratings for each sentence were then used as a gold standard against which output of a model can be compared, and the results showed that NLP models have roughly comparable performance levels to human subjects' gold standard thereby offering a how-plausible explanation of human language processing. They conclude with the following remark:

> "If one is to sustain a categorical theory of grammatical competence by attributing gradience to performance and processing, it is necessary to formulate a precise, integrated account of how these two mechanisms interact [...] The relative contributions of competence and of performance devices must be testable [...] If this is not the case, then competence retreats to the status of an inaccessible theoretical posit [...]" (2016: 33).

The integrative account of competence and performance that Lau, Clark and Lappin plea for is out of the question for anyone endorsing TGG. The whole point of distinguishing competence from performance is to discard the influence of any extra-linguistic factors and devices that would be merely noise in the otherwise abstract, domain-specific linguistic realm. This is usually backed up by holding grammatical competence cognitively autonomous from other sensory-motor and content-based processing. Gabe Dupre (2021b) denies any relevance of DL models of NLP for theoretical linguistics on the basis of this distinction: actual linguistic usage in training datasets is often uninformative and hardly reflective of "linguistic-specific innate structures" and these structures are dependent on internal organization. In his view, any DL models operate on surface observable utterances, that may be more or less acceptable, but the patterns thusly produced within models do not correspond to rule-governed grammatical competence and cannot say much about it. In the lingo of philosophy of language, Dupre highlights that DL models aim at identifying function-in-extension, i.e., the finite set of utterances encompassing probability distributions; while theoretical linguistics strives to handle function-in-intension, i.e., content and structure of presupposed linguistic entities.

Pace, empirically enlightened linguists and empiricist philosophers. This is, after all, something you have already heard many times. Dupre's stream of thought can be easily filled in with previous points moderate rationalists like Gary Marcus advanced, for instance, that structure-sensitivity of linguistic representations must be a built-in feature of a model for NLP in the form of hierarchical tree structures. Both Dupre and Marcus would agree that DL models cannot simulate linguistic competence faithfully enough and both will be claiming that from their TGG corner that assumes competence vs. performance distinction as preliminary step in the argumentation against connectionism. In words of Christiansen and Chater,

> "[S]uch traditional distinctions aim to split apparently related issues into two: that is, to suggest that superficial indications apart, Topic A and Topic B require different kinds of answers [...] and hence that the study of Topic A and Topic B should proceed in relative, or even complete, isolation from one other" (2016: 228-229).

And as results of Lau, Clark, and Lappin suggest, there is little empirically grounded reason to opt for fragmentation between unobservable theoretical entity and our everyday linguistic experience rather than testable integration. In what follows, I try to turn the tables in favor of the DL models of NLP: I distinguish syntactic from semantic competence to show that TGG endorses an overly narrow conception of linguistic competence, and build the case for integrative, embodied, UBT-friendly account of linguistic capabilities. Finally, and most controversially, I propose to put the kibosh on competence vs. performance distinction.

4.2. Syntactic Competence

Syntactic linguistic competence amounts to being able to distinguish sentences from other linguistic items. The origins of this competence were seen in recursion, which can be construed in connectionist terms judging by the results of Sect. 3.2. Nonetheless, here we focus on syntactic processing embedded in the *theoretical entity*, viz., syntactic competence. In TGG framework, syntactic competence is assumed to be cognitively autonomous. This essentially means that rules and structures involved in sentence formation are domain-specific and encapsulated from other forms of language processing such as semantic processing, as well as from the other forms of cognitive processes, such as sensorimotor processing. Let me formulate the view of syntactic competence against which I aim to argue in the rest of the Sect.

> (**SyntComp$_F$**) Syntactic competence is constituted by *symbolic or syntax-sensitive representations* and *discrete rules* allowing for the manipulation of representations. Additionally, syntactic competence is neither linked to semantic processing, but is rather *independent*, *innate*, and *domain-specific.* For this reason, accounts *à la* TGG are able to account for this uniquely human capability since TGG provides enough evidence in the form of the Poverty of stimulus argument.

It follows from (**SyntComp$_F$**) that DL models specialized for NLP are unable to account for syntactic competence. Usually, the key evidence that is taken to support (**SyntComp$_F$**) is the poverty of stimulus argument. I disagree with each italicized word in (**SyntComp$_F$**) and

through tackling the reasons for disagreement, I argue in favor of the **Main hypothesis** and **Auxiliary hypothesis A**. I assess the syntactic capabilities of various DL models through two linguistic phenomena, namely long-distance dependencies and question formation, because generativists consider these as invariant properties of natural languages which figure prominently in the most recent formulations of the Poverty of stimulus argument (Berwick et al. 2011). I follow the three crucial claims (1)-(3) that I have previously defended for linguistic competence in general, albeit now applied for syntactic competence specifically. Thus, I aim to show that syntactic competence can be *examined*, *explained* and *simulated faithfully enough* with DL models. Along the way, I also challenge the Poverty of stimulus argument, and show that accepting (**SyntComp$_F$**) means to endorse a vacuous theoretical entity—an idealized conception of capability that is neither shared by humans nor machines.

Long-distance agreement in linguistics refers to syntactic dependence or agreement between noun/subject and verb/predicate as well as other sentential elements. As Linzen & Baroni (2021) point out, this syntactic phenomenon has been singled out by generativists as being prototype of "structures, not strings" manifesto which essentially states that these syntactic dependencies are indicative of an abstract, hierarchical, and tree-like structure rather than mere linear sequence of words. Should it be the case that DL models for NLP can handle long-distance agreement, that would indicate that ANNs can at minimum learn to approximate syntactic structure, or at maximum to faithfully simulate it. Linzen, Dupoux, & Goldberg (2016) explored whether LSTM can handle long-distance agreement via the number prediction task. A model, being first of its kind, is exposed to English sentence prefixes extracted from a natural Wikipedia corpus, rather than from an artificially constructed set as generativists usually do. The model is then tasked to predict plural or singular form of verbs following the prefixes, for instance the sentence "Alluvial soils carried in the floodwaters_____ nutrients to the floodplains" (Linzen, Dupoux, & Goldberg 2016: 523). In this way, LSTM is checked for producing grammaticality judgments, and *ipso facto*, for its syntactic capabilities. The accuracy of the LSTM implemented in the model was astonishing: Linzen and colleagues report a range of 99% to 82% accuracy, depending on the number of intervening nouns, such as "floodwaters" in the sentence above. These intervening nouns or noun phrases are also called attractors.[71] Agreement attraction errors are tricky for LSTM performance given that they arise from over relying on attractors arising from exposure to data arranged in linear sequence instead of processing syntactic number and syntactic subjecthood. Naturally, the more attractors occur between noun/subject and verb/predicate, the error rate increases, especially if the ANN was fed only with the noun/subject and without any additional syntactic cues. Nonetheless, LSTM generally managed to handle both regular and rare distances, from 1 to up to 15 words. It is worth noting, however, that Linzen and colleagues did give explicit feedback about correct verb number, i.e., relied on supervised training of the ANN.

[71] Attractors can include sounds, forms, or grammatical structures that are being selected when language contact and everyday communicative situations make it possible for speakers to choose between competing syntactic forms (Nichols 2018). The term is actually borrowed from the complex systems theory where it designates a state that is easier to acquire or retain than to lose. In most cases, attractors ease the load of cognitive processing and contribute to faster understanding albeit at the expense of correct grammatical structure.

To check the robustness of their results and see which tool seems more promising for psycholinguistics, they also designed a language model implementing an RNN trained in an unsupervised manner. When the RNN was exposed to each word from the input, a fully connected dense layer formed along with a layer that spans the entire vocabulary, and the task at hand was to assign probabilities to verb forms (Linzen, Dupoux, and Goldberg 2016: 528). Researchers report that language model made eight times as many errors than the original model with LSTMs, and increasing in scale by adhering to Google's proto-LLM has not changed the overall poor performance once the attractors begin to mess with processing (Linzen, Dupoux, and Goldberg 2016: 530). The two important takeaway messages follow from this inaugurating research on syntactic capabilities of DL models. First, Linzen and colleagues' approach showed that it is feasible to develop models that could exhibit more advanced generalization skills by gradually introducing them to grammatically challenging training sentences. Second, LSTM in the original model did prove to be sensitive to syntactic number and subjecthood, and even the kind of errors that were made in relation to the number of attractors are qualitatively similar to documented attractor agreement errors in humans, especially their acceptability judgements (Linzen, Dupoux, and Goldberg 2016: 532).

This echoes Lau, Clark, and Lappin's point that neither grammaticality nor acceptability judgments are categorical, meaning that competence mirrors the stochasticity of performance. This further suggests that a performance model may be constructed from a competence model and vice versa given the additional assumptions and data from psycholinguistics (*cf*. Linzen 2019: e106). Moreover, as Tal Linzen explicitly states: "[D]ivergences between human behavior and the behavior of a neural network can suggest ways in which the architecture can be changed to better model humans" (2019: e105). Thus, again in line with being a research program in development, connectionism is better off with differences rather than similarities since differences are sometimes more likely to hit the nail on the head. Linzen and colleagues' models suffer from three limitations that ought to be surpassed within such research program in order to arrive at how-plausibly explanation of syntactic agreement: diverse architectures should be tested against the same task to check for robustness of results, to validate emergent syntactic ability, human and ANNs qualitative and quantitative error patterns should be compared and given that syntactic supervision was still needed for better performance, and without further improvement in this regard, empiricists are quite put on the spot. Bernardy & Lappin (2017) grapple with the first limitation, Linzen & Leonard (2018) with the second, while Gulordava et al. (2018) defend the empiricist cause.

Interestingly, Bernardy & Lappin (2017: 3) claim that their primary epistemic goal is the very process of exploration, i.e., toying with the model. Hence, they experiment with several computational architectures (LSTMs, regular RNNs, GRU-based RNNs, CNNs) and tweak some of the parameters (e.g., ratio of training to testing with respect to the size of the corpus in English, memory size, vocabulary size, word embedding dimension size, etc.) as well as hyperparameters to see how the ANNs behave so that they could compare the results to Linzen, Dupoux, and Goldberg (2016). The researchers were convinced that the efficient learning of syntactic agreement within the number prediction task would depend on feeding the ANNs with lexically impoverished sequences since no semantic cues would camouflage purely syntactic representations such as syntactic number and subjecthood. Their results indicate that LSTMs, RNNs, and GRU-based RNNs all perform comparably

well on the number prediction task and achieve significantly better accuracy than CNNs (Bernardy & Lappin 2017: 8). Their success in handling multiple attractors scales with the size of training dataset. However, surprisingly, the size of vocabulary and word embeddings also affect the performance: the increase in lexical semantic cues actually improves learning of syntactic dependencies contrary to researchers' initial hypothesis.

Two key insights are linked to Bernardy & Lappin's results. First, DL models for NLP *do* capture abstract syntactic structure given that their successful performance on the number prediction task does not hinge on specific architectural features. Thus, ANNs are both useful tools for exploring constraints on syntactic phenomena such as long-distance agreement and show that syntactic competence can be examined by DL models for NLP given that the results on task for testing their syntactic capabilities are robust. Second, as for the independent evidence in favor of the fact that these models learn more efficiently syntactic patterns more readily if the richness of word embeddings is introduced, this is indeed supported on the level of neuronal mechanisms. Blank et al. (2016) report that neuroimaging studies of individual subjects show the distributed nature of syntactic processing thereby pointing out that syntax is inseparable from broader lexical and semantic processing. In other words, syntactic complexity is not localized, i.e., confined to a particular brain region of a language system, but can be found throughout the entire system (Blank et al. 2016: 314). This can be used for advancing the argument that the assumption of the cognitive autonomy of syntax has no neural or implementational underpinning. Moreover, given that results of Bernardy & Lappin suggest that subsymbolic or subpersonal level (recall Sect. **2.2.** and **2.3.**, also fn. 59) on which the syntactic patterns are captured by ANNs clearly shows that the richness of semantic cues influences syntactic processing, it can be inferred that this assumption has no plausible algorithmic underpinning either. Nonetheless, a path towards how-plausible explanation should include calibration with human data given that generativists assume that human syntactic competence encodes innate explicit structure-sensitive representations, e.g., hierarchical tree structures. Besides, Bernardy & Lappin's DL models are also trained via supervised learning and do not check for morphologically rich languages, i.e., languages in which number, gender, and person are morphologically realized on verbs, which means that performance on the number prediction task would be assessed across three dimensions rather than only one as in the case of English corpus.

Linzen & Leonard (2018) reveal that RNNs and humans have qualitatively similar error patterns. Thus, errors are more likely to occur when attractors are present in dataset, especially when the subject of a sentence is singular and the attractor is plural, attractors are more problematic in prepositional phrases than relative clauses (Linzen & Leonard 2018: 694). However, RNNs showed increased error rates in processing relative clauses as opposed to humans, which indicates that they are using divergent heuristics. Relative clauses, such as "The social reforms in the Roman Republic that Tiberius and Gaius Gracchus set in motion...," are easier to process because the subject is insulated from the words inside the clause. Linzen and Leonard (2018: 696) realized that RNNs rely on the (wrong) heuristics that relative clauses are short, so they usually stop processing before arriving at attractor, unless the clause is complex. Their final assessment is not encouraging for empiricists: researchers should either supply models with syntactic annotations or turn to manually specified priors and biases. Truth be told, Linzen and Leonard's final assessment stems from their engineering, entrepreneurial spirit rather than from taking

sides in the debate between empiricists and nativists. In any case, were they to choose sides, they would probably align with Joe Pater's ecumenist proposition (*cf*. Linzen 2019). Nonetheless, the ramifications of their study are of greater relevance for the time being: a path towards how-plausible explanations of syntactic agreement must address the peculiar way in which RNNs choose "surface" heuristics.

In my view, this should not strike the reader as a mysterious ramification that is indicative of the ultimate implausibility of NLP in an RNN. Take, for instance, Marvin & Linzen (2018) study which shows that LSTMs *only* failed to process syntactic dependencies in sentences that rarely occur in the corpus, e.g., more complex examples containing nested long-distance agreement. In other words, across four studies (Linzen, Dupoux, & Goldberg 2016, Bernardy & Lappin 2017, Linzen & Leonard 2018, Marvin & Linzen 2018), ANNs consistently showed successful performance on the number prediction task, i.e., the capability to produce abstract syntactic patterns without encoded syntactic inductive bias or prior, albeit within a supervised learning setting. The peculiar way of handling syntactically complex sentences aligns with their training dataset. Being stochastic models, thereby following the good old GIGO (short for garbage in, garbage out) adage, RNNs and LSTMs in fact need more syntactically diverse corpus to account for rare syntactic phenomena—as do we. As per Zipf's (1949) Principle of Least Effort, speakers tend to simplify language and, in the course, they become sloppy and lean on frequent and irregular constructions. In that sense, LSTMs having problems with rare and complex syntactic dependencies is actually more plausible than processing them flawlessly. Becoming better in syntactic processing goes hand in hand with being exposed to more complex patterns that bypass the Principle of Least Effort that guides our mundane communication. Hence, for obtaining how-plausible explanations, the quality of data seems more relevant than mere quantity. Qualitative error patterns among humans and ANNs are also consistent across studies. Thus, humans are also notoriously bad at processing the nested long-distance agreement (Grodner & Gibson 2005, *cf*. Lakretz, Dehaene, & King 2020), which aligns with Marvin & Linzen's findings.

As for the "superficial" heuristics in the case of syntactic dependencies within relative clauses, the solution presented itself with the resurgence of transformer architectures crunching textual data across the internet. Unlike RNNs, transformers do not process words incrementally over time but rely solely on attention mechanisms, or attention heads, that provide the ANN with access to words from all previous contexts. This means that from the start transformer cannot use heuristics such as short relative clause given that it would be processed in its entirety at t_1 and compared to earlier instances of tokens constituting the clause at t_2. Using the same stimuli and slightly changed experimental protocol to fit transformer architecture, Goldberg (2019) evaluates the syntactic capabilities of one of the first LLMs, namely BERT. BERT is pre-trained masked LLM, which means that the model first learns to fill in random gaps in the unlabeled dataset to obtain relevant vector representations and only then is exposed to regular training and validation. Having been exposed to some 3.3 billion textual data from Wikipedia and unpublished books within Toronto BookCorpus, BERT served as a baseline for NLP tasks and benchmarks (the so-called BERTology). Thus, in the number prediction task, transformer-based BERT grasps the entire sentences at once (minus verb, of course) instead of only processing prefixes like sequential LSTMs and RNNs. BERT scored high performance numbers even in sentences with as many as four attractors and outperformed recurrent-based ANNs (Goldberg 2019:

3). Rogers, Kovaleva, and Rumshisky (2020) further review BERT's syntactic knowledge and report that syntactic structure is not directly encoded in attention heads, but that vector representations are hierarchical rather than linear. This suggests that LLMs, such as BERT, do vindicate empiricist ambitions to some extent. However, these researchers also note that BERT is, unfortunately, insensitive to distorted input and has trouble processing another crucial syntactic phenomenon I want to address here, namely question formation. Thus, BERT indeed enhances the plausibility and feasibility of DL models with respect to processing long-distance dependencies, which adds up to the explanatory prospects of these models, but it can hardly be deemed a successful simulation of human syntactic competence. More importantly, we have not tackled the Poverty of stimulus argument yet, and this argument is the first barricade standing between how-actually explanations of syntactic competence and DL models. The argument amounts to claim that language acquisition, most notably syntactic structures, cannot be derived from input alone since the relevant data is either not present or not frequent enough in the input so it must be innate.

Returning briefly to the syntactic dependencies, Gulordava et al. (2018) wanted to examine to what extent LSTMs track abstract syntactic structure—as shown in the studies of Linzen, Dupoux, & Goldberg (2016), Linzen and Leonard (2018), and Marvin and Linzen (2018)—without being given explicit instruction to focus on long-distance agreement. Following Chomsky's (1957) idea that we are able to process meaningless sentences like "The colorless green ideas sleep furiously" thanks to the structure-sensitive and innate representations, Gulordava and colleagues test DL model on regular and nonsensical sentences in English and morphologically rich languages such as Italian, Hebrew, and Russian. Recall that morphological rich languages were singled out in the previously mentioned studies as particularly demanding for ANNs when it comes to performing the number prediction task. However, LSTMs on average performed quite well for both original and nonsensical sentences on the condition that the language is morphologically richer rather than poor as English (Gulordava et al. 2018: 1199-1200). While validating results through comparison with human subjects' performance, Gulordava and colleagues noticed that easiest and hardest examples were the same for both humans and ANNs. Finally, the researchers conclude that their results mirror human experimental studies showing that morphologically rich languages correlate with fewer errors linked to the attractors given that there is less ambiguity thanks to inflections for gender, number, and person. This indicates that input obviously has enough initial information to trigger syntactic learning since LSTMs were not given any prior knowledge about long-distance dependencies, which runs counter to the Poverty of stimulus argument. Moreover, LSTMs did not merely memorize morphosyntactic linear sequence, rather they engaged in active syntactic processing.[72]

[72] Lakretz et al. (2019) did a follow-up study on Gulordava et al. (2018) to check their results by analyzing the emergence of syntactic constituents and number in LSTMs. They report that LSTMs allocate a scarce number of units to handle long-distance information and that these units are connected to a subnetwork of units that are sensitive to syntactic constituency. The subnetwork can be examined separately from the rest of the architecture and can process grammatical structure. This means that LSTMs do not use surface-bound heuristics but rather exhibit functioning as per underlying mechanism for syntactic competence. This goes to show that the performance of ANNs yields insights for competence despite the prevailing competence vs. performance distinction—I will return to this point in the next Sect.

Now, to strengthen the case against TGG and in favor of moderate empiricism, I turn to question formation and show that ANNs are capable of hierarchical generalization without encoded hierarchical structures or rules by drawing on McCoy, Frank, & Linzen (2018). The researchers focus on a particular question formation, namely subject-auxiliary formation that Chomsky (1957, 1980) stipulated to be quite rare in child-directed speech. This syntactic phenomenon arises when one takes sentence "Spartacus can lead an uprising of slaves" and transforms it into the question "Can Spartacus lead the uprising of slaves?". The hierarchical rule that allegedly runs the processing predicts that the right thing to do would be to move the main verb's auxiliary to the front of the sentence, i.e., to place "can" before "Spartacus". The things get messy when we add central embedding as in the question "Can Spartacus who will be executed lead the uprising of slaves?" since there is a tension between two verbs "can" and "will". Linear processing would predict the ungrammatical version of the question, and this is precisely what generativists would accuse DL models of. However, McCoy, Frank, & Linzen (2018) check the performance of RNNs, LSTMs and GRUs with attention heads on this task and report that ANNs with attention heads proved most successful in predicting the auxiliary because they learned to ignore linear information thanks to the attention mechanisms. By having only hierarchical cues in the input sentences, these ANNs managed to produce hierarchical generalization, and adding central embeddings only increased the likelihood that ANNs would behave hierarchically. In other words, neither domain-specific knowledge nor structure-sensitive representations were needed for the processing of this particular syntactic phenomenon. Or, to put it succinctly, the barricade has been jumped over.[73] This, of course, does not mean that we have found ourselves in Eldorado of how-actually explanations.

A Doubting Thomas around the corner could say that both cases against the Poverty of stimulus argument are based on DL models that are either using small fragment of English (McCoy, Frank, & Linzen 2018) or probably copying from training data since it is nowhere near clear whether they can express linguistic novelty (Gulordava et al. 2018). Children are not stochastic parrots; their innate syntactic constraints enable them to generate grammatical and coherent sentences that are distinct from anything they heard before. I turn to another transformer-based LLM to reinforce my argumentation— this time GPT-2 comes to rescue. McCoy et al. (2021) evaluate the novelty of generated text in English by GPT-2 via cleverly dubbed set of analysis—RAVEN, which includes, *inter alia* inflectional morphology and syntactic agreement.[74] However, researchers are not interested in prompting the model to generate nonsense, but rather high-quality text measured by a pointwise duplication score, which tracks the extent to which generated sequences

[73] I have focused here on computational means for debunking the poverty of stimulus argument since I am interested in whether DL models for NLP and LLMs can account for syntactic competence. However, it is worth noting that there is a growing discontent with the allegedly intact authority of the poverty of stimulus argument among linguists. For instance, Ewa Dabrowska, a leading cognitive linguist summarizes all empirical shortcomings of crucial aspects of TGG related to the idea of universal grammar in her 2015 paper, including the debunking of the poverty of stimulus argument. As she rightly remarks: "Strikingly, most expositions of the poverty of stimulus argument in the literature do not take trouble to establish the truth of the premises: it is simply assumed" (2015: 14).

[74] The acronym refers to Edgar Allan Poe's poem The Raven, in which it is unclear whether raven uttering "Nevermore" is simply repeating something previously heard or conveys a genuine message to the poet. Here is a hint: counting Bender et al.'s famous metaphor of "stochastic parrots", only black-capped chickadees' vocalization repertoire remains unused in virtue of being an analogue to DL models' linguistic capabilities.

duplicate previously seen contexts. Findings are particularly promising. GPT-2 shows a high degree of novelty in the domain of syntax given that the majority of generated sentences had syntactic structure that was not present in the training dataset (McCoy et al. 2021: 6–7). Additionally, in the domain of morphology, 96% of newly generated words are morphologically well-formed, and they fit the syntactic context in 94% of cases. McCoy and colleagues (2021: 9) also notice that compositional generalization is feasible at least in morphology: GPT-2 combines familiar prefixes and suffixes with never encountered word stems. Closely related to my remark about the growing need for more diverse rather than larger corpora, McCoy and colleagues also notice that researchers ought to revise their intuitions regarding what can be expected to occur naturally in corpora. Thus, even though generativists assumed that regular past tense forms of irregular verbs are virtually absent from children's experience and bootstrapped by universal grammar which impedes such linguistic abominations, training dataset of GPT-2 includes as much as 92 instances of this abomination (McCoy et al. 2021: 11), thereby suggesting that acceptability/grammaticality judgments can stretch *really* far.

What alternative am I offering now that TGG is out of picture? I turn to the Construction Grammar Approach (CAP), pioneered by eminent American cognitive linguist, Ronald Langacker (1987, 2008), which was recently coupled with UBT (Barlow & Kemmer 2000, Tomasello 2003).[75] CAP presumes that language is constituted by constructions rather than structures, whereas constructions are understood as form-meaning pairings that were picked out by language usage and experience of specific language speakers.[76] This is where a link is forged between CAP and UBT: UBT also emphasizes actual language usage as being fundamental for language learning and shaping of the linguistic competence and regards the more frequent linguistic patterns as more salient (Langacker 2008: 220). Let me apply the coupled frameworks to the issue of syntactic competence. First, and most importantly, CAP is based on the non-autonomy of syntax and the non-modular view of language (Langacker 2008: 14), also endorsed by UBT. In Langacker's terms, syntactic representations are *patterns* of grammatical constructions and exhibit significant individual variability. Or, as Dabrowska puts it: "Languages are shot through with patterns [...] Languages are also shot through with idiosyncrasies: constructional idioms, lexical items which do not easily fit into any grammatical class, irregular morphology" (2015: 15). Syntactic constituency does not come from the structure-sensitive representations that ought to mirror hierarchical tree branching as generativists assume. Instead, it is an emergent feature of mutually constraining morphological, semantic, and pragmatic factors that shape syntactic patterns, which stem from domain-general mechanisms and real-world experience (Goldberg 2003: 219–220, Langacker 2008: Subsect. 8.1.2.). What a generativist would call a low-level template and performance-related phenomenon, as opposed to deep structures representative of competence, a constructionist would call learned constructive schema representative of competence embedded in the external setting. In TGG, linguistic universals, like subject-auxiliary formation, are viewed as manifestations of innate universal grammar. On the other hand,

[75] Both the Construction Grammar Approach and UBT belong to a wider anti-TGG front, namely cognitive linguistics, a broad linguistic discipline uniting all outcasts from the generativist tradition, including proponents of embodied cognition and mind such as Lakoff & Johnson (1999). For a brief history and general overview of cognitive linguistics see (Langacker 2008: Ch. 1).

[76] *Nota bene*: Constructions here should be taken as both linguistic items and cognitively realistic patterns that can be detected and experimentally tackled.

in CAP, this would be seen as a regular inductive generalization based on observable features of a specific language. A generativist divides competence and performance, cuts syntax from semantics, while a constructivist integrates competence and performance and builds a syntax-semantics interface. A generativist sticks to the idealized and unsullied image of language, while a constructivist is faithful to the actual usage of language conceived as de-idealized and evolving capacity that reflects the linguistic needs of particular communities.

CAP and UBT can both subtend post-connectionist modeling of NLP *qua* fitting *and* illuminative theoretical frameworks given the joint *credo in unum* pattern-governed cognition scaffolded by multiple interacting constraints and external environmental factors. To prove that this is not a mere philosophical suggestion, but an empirically testable proposition consider a recent study by Madabushi et al. (2020). They show that BERT has, in fact, access to constructional information by creating probes requiring BERT to predict if sentences are instances of the same construction. After brief training on 500 sentences, BERT arrived at 90% accuracy of predicting sentences. Researchers also check whether fine-tuning and manually specifying priors changes BERT's efficacity and report that the answer is negative. Madabushi and colleagues conclude with quite inspirational tone:

> "The impact of this observation is potentially far reaching as it not only further shows the capabilities of [DL] methods, but also shows that information that is typically called constructional can be learned from exposure to lexico-semantic information. This is expected given [that] words and constructions constrain each other mutually" (2020: 10).

A path towards how-actually explanations of syntactic competence would have to include a tighter connection to CAP and UBT. In other words, LLMs syntactic powers should be assessed from the perspective of these two frameworks rather than TGG in similar vein as Madabushi et al. (2020), only for a wide range of syntactic phenomena. This means that we need new experimental settings, new benchmarks, and an additional batch of human data from cognitive linguistics.

Finally, let me introduce a rivalrous view to (**SyntComp$_F$**), namely (**SyntComp$_T$**), which is backed up by previous considerations in this Sect., and renders a transparently formulated defense of the **Auxiliary hypothesis A**.

> (**SyntComp$_T$**) Syntactic competence is constituted by *stochastic patterns* resulting from parallelly distributed vector representations. Additionally, syntactic competence is not isolated but intertwined with semantic and lexical processing, and it is neither innate nor domain-specific. Instead, syntactic competence is a *learned* capability emerging from *domain-general mechanisms*. For this reason, TGG is unable to account for this capability since this framework does not grant convincing evidence, given the failure of the Poverty of stimulus argument; rather, CAP coupled with UBT is more adequate option.

The main problem with (**SyntComp$_F$**) was that it advances a theoretical entity that is vacuous and isolated—i.e., neurally implausible and dissected from semantic and morphological processing, which goes against the available human behavioral and psycholinguistic data. In other words, it is a theoretical entity that remains theoretical. Nonetheless, a novel account of linguistic competence that I am advancing in the dissertation has another important aspect, namely semantic competence. Hence, in addition to (**SyntComp$_T$**), a (**SemComp$_T$**) should be introduced and defended in similar vein so that I could make a compelling case for the new empiricist dogma in the domain of NLP.

4.3. Semantic Competence

The case for semantic competence in connectionism was doomed at the very start, as Horsemen of Apocalypse, Jerry Fodor and Zenon Pylyshyn predicted back in 1988. At least in theory, syntactic competence was thought to be reachable if (and only if) generativists and connectionists would meet each other halfway, i.e., by implementing hierarchical tree structures as priors of NLP models and acknowledging the innateness of the narrow language faculty responsible for grammatical knowledge. Semantic competence, however, was deemed unreachable *in principle* in a computational model based on ANNs of any sort: parallelly distributed representations cannot account for semantic content since they are not structure sensitive, which further implies that compositionality, systematicity, and productivity cannot be met within such a model. Recall Gabe Dupre's philosophical argot: we can grant DL models the capacity to handle function-in-extension, given their computing of probability distribution, albeit this does not hold for the function-in-intension. The other unfortunate circumstance for semantics is its Ianus-faced nature shifting between formal and psychological in the prevailingly quite myopic publications of philosophers, linguists, cognitive scientists, and AI researchers. Formally, any theory of natural language semantics must account for its compositionality. Psychologically, semantic processing must give credit to systematicity and productivity. Formally, the meaning of "meaning" can encompass as many interpretations as there are analytic philosophers discussing it. Psychologically, the meaning of "meaning" can encompass as many concepts as there are computational models and experimental studies. Conveniently, traditional symbolic cognitive science seemed to unite formal and psychological accounts of semantics since the postulated representations had formal properties, viz., combinatorial structure, shared between language and thought, thereby offering the ultimate internalist *Bauplan* (recall Sects. **1.3.**, **2.1.**, and **2.2.**).

However, as Potts (2019: e115) remarks in response to Pater's ecumenist position, a DL-based semantic theory is specific precisely because it has replaced logic as "the most-used toolkit in the field" given that each function-in-intension has been swapped for vectors and matrices. To show that connectionism can account for both formal and psychological aspects of semantics, it is necessary to build a case that DL models for NLP can handle compositionality and systematicity, albeit in their peculiar and novel manner. This peculiar and novel manner concerns vector representations and comes with significant change in metaphysical realm as to how the nature of semantic processing should be understood. In this Sect., I finalize the defense of the **Main hypothesis** and **Auxiliary hypothesis B** stating that DL models of semantic processing should be intertwined with the embodied approaches to cognition, which essentially means that they have to be grounded in multimodal data so that their output forms a feedback loop with body and environment. This would enable the formation of unified theoretical framework for syntactic and semantic competence through linking CAP, UBT, and core tenets of embodied cognition. These are metaphysically innocuous and mutually compatible takes on language and cognition that can be tested through post-connectionist models and, therefore, they can contribute to the inauguration of connectionist domain-general multimodal cognitive architecture advanced by moderate empiricists. As in the case of syntactic competence, I follow the claims (1)-(3) into which I have divided the core of the **Main Hypothesis** in Sect. **4.1.**, i.e., that DL models of NLP and/or LLMs serve as tools for examining semantic

competence, offer at least how-plausible explanations, and simulate faithfully enough human semantic processing. Nonetheless, this will not proceed neatly and methodically as in the case of syntactic competence. While DL models of syntactic processing can target how-plausible explanations without spurring any particular controversy, the connectionist case for semantic processing is much less reassuring despite the remarkable success of LLMs. Thus, the ultimate evaluation of the **Main hypothesis** will hinge on what "simulating *X* faithfully enough" really refers to in this domain. Besides, the most severe limitations of these models concern the systematicity problem and are lethal for each of the claims—this problem hits too close to home since I strive to defend connectionism *qua* autonomous theory of human cognitive architecture.

As philosophers almost always do, I start with foundations: success (or, alas, a failure) in dealing with the principle of compositionality is without any doubt a candidate for building foundations of the connectionist theory of semantics. Recall from Sect. **2.2.** that compositionality concerns semantic derivation: the meaning of a complex linguistic expression is determined by the meaning of constitutive simpler linguistic expressions and the manner in which these expressions are combined. Nefdt (2020: 55) remarks that the principle of compositionality presumes that constituent parts must be *meaningful* and that we must be able to *identify* such parts, and then goes to distinguish between three types of compositionality that can be found in literature, namely Process Compositionality, State Compositionality, and Outcome Compositionality. Process Compositionality tracks rule-to-rule mapping between constitutent parts to form a complex expression. This sort of incremental compositionality is, for instance, assumed in Pylyshyn & Fodor (1988) when accusing connectionism for not being able to handle compositionality. State Compositionality has to do with the decomposable constituent parts—since they could have been decomposed, they must exhibit compositionality. Nonetheless, as Nefdt emphasizes, these two types of compositionality are independent from each other. This means that it is quite possible for a compositional process to generate a non-compositional state, i.e., we do not have to be in the epistemic position to identify all constituents that were used for composing complex linguistic expression (Nefdt 2020: 57). Finally, Outcome Compositionality is purely functional and conveys the idea that given a certain input, the output will be compositional. Again, this type of compositionality does not entail State Compositionality since it can be evaluated without having to arrive at the end of the expression and store it as whole. In this case, not all parts need to be meaningful, because the computation will take into account all parts before resulting in the output. The issue arising in DL modeling stems from confusing these types of compositionality and using success or failure in exhibiting Output Compositionality to account for either Process or State Compositionality.

Consider Lake and Baroni's paper from 2018 where they argued that ANNs—specifically SRNs, RNNs, LSTMs, and GRUs—are "still not systematic after all these years", which was then readily exploited as a support for rationalist criticisms of DL models as in Marcus (2018a). The "still not systematic after all these years" argument proved useful to ignore virtually all trench raids in the 21st century that were gaining territory for connectionism (e.g., Frank, Haselager, & van Rooij 2009, Frank 2014). Lake and Baroni (2018:

2) use the so-called SCAN dataset that translates commands in simplified English language into sequences of action, which essentially amounts to supervised semantic parsing. Commands could be decoded compositionally if ANNs acquired the right interpretation function, i.e., function-in-intension, which would allow for constructing complex previously unseen commands from constituent ones in the training portion of dataset. The researchers claim that systematic differences between training and test sentences have a disastrous influence on the behavior of these ANNs which were tasked to perform zero-shot generalization. When ANNs were tasked to bootstrap to commands demanding longer action sequences which were not present in the training dataset, their performance on average achieved 13.8% accuracy even though they manage to get right even the longest sequences—albeit only if they are somewhat similar to training sentences. The additional experiment included composition across primitive commands: ANNs were exposed to simple and compositional commands and during test phase were tasked to produce all compositional commands from one simple command. The results suggested that even when ANNs showed signs of systematic compositionality, this seemed far from human processing since, for example, for command "jump" barely any compositional command was obtained, as opposed to the case with commands "turn left" or "walk". As Lake & Baroni remark (2018: 7) ANNs seem to fail in adhering to categorical, i.e., all-or-nothing, behavior and this makes them very "unhuman" since they, apparently, lack compositionality that would have to be manually specified by introducing symbolic representations into ANNs (how original and never heard off).

My issue with this piece of evidence is that it conflates Outcome Compositionality with Process Compositionality and as a solution proposes State Compositionality. Let me unpack this. Lake and Baroni conclude that ANNs fail in exhibiting compositionality *qua* process based on their task performance, i.e., concrete outputs. However, they have not analyzed in what sense the mapping from constituents (viz., primitive commands) to complex whole (viz., compositional commands) does not get off the ground; rather, they concluded that based on the input-output relation within a model. Their proposal then includes tweaking of representations, which suggests that they seem to assume that *states* are responsible for exhibiting compositionality. However, as I have described above following Nefdt (2020), Outcome Compositionality, the only type of compositionality subtended by their model, does not entail State Compositionality nor State Compositionality would add up to the Outcome Compositionality or Process Compositionality since the State Compositionality need not to be related to it. The second more general issue is that the researchers interchangeably used terms "generalization" and "compositionality". In ML literature "generalization" refers to cases when models make accurate predictions about data that were not encountered during training and creatively process or produce such data (see e.g., Hadley 1994). As Potts (2019: e120) correctly claims, generalization is something we can measure using benchmarks and other quantitative metrics, while compositionality is a highly restrictive formal strategy which could hinder processing of useful information *a priori.* In other words, measuring generalization is not the same thing as evaluating compositionality, and in any case, adhering to a particular quantitative metrics says nothing about the type of compositionality that could be evaluated in principle.

Moreover, Groenendijk & Stokhof (2005) argued a couple of years before mushrooming of DL models that compositionality can have either completely different role or no role whatsoever in mathematical frameworks that do not rely on the formal language of classical logic. This issue could not have occurred earlier given that GOFAI was based on formal language of classical logic and so the etalon was transferred to the 21st century similarly as requirements for full transparency with respect to inner machinery (recall Subsect. **Thinking outside the (Black) Box**). Nefdt (2020) also claims that compositionality that was at the forefront of logic and philosophy of language cannot be accessible in ANNs given that their inherent opacity stops us from identifying meaningful parts of the compositional structure. In his terms, therefore, neither State Opacity nor Process Opacity can be fully discernible. I do not entirely agree with Nefdt since XAI approach targets what could be described as meaningful parts of the input to shed light on what is going to be processed "inside" the model.

Lake & Baroni, as well as Nefdt, make quite an important point but neither of them draws the right consequences from it: compositionality, that we philosophers know of, indeed cannot be obtained in DL models—as it is the case with traditional notion of explanation, but the alleged "unhuman" lack of compositionality suggests that this is an ill-suited (and ambiguous) etalon from the very start. As Baroni admits in a later publication:

> "Classic and modern criticism of [ANNs] emphasizes the aspects of human language that are best characterized by clear-cut, algebraic rules. Language, however, is also host to plenty of productive phenomena that obey less systematic, fuzzier laws..." (2019: 10–11).

This is precisely what I will be arguing in the rest of the Sect., albeit with more evidence as opposed to Baroni's mere remark in passing. I propose to zoom in on the representations within DL model to see what kind of systematicity can be expected of ANNs as well as to sketch fitting foundations for the connectionist theory of semantics. Dasgupta et al. (2020) analyze word embeddings of InferSent, the DL model that implements encoder-decoder architecture with recurrent and convolutional units, and then test it on the natural language inference task. InferSent thus should group 550 thousand sentences arranged in an argument, i.e., premises and conclusion, as entailments, negations, or neutral by grasping semantic relations between the sentences. The labeled dataset was obtained via crowd sourcing similarly like Lau, Clark, & Lapin (2016), and, therefore, model was trained through supervised learning algorithms. Contrary to Lake & Baroni (2018), the researchers did not just assume failure in the Process Compositionality based on Outcome Compositionality and demanded a rationalist intervention akin to State Compositionality; rather, they studied incorrect grouping of sentences to unravel what mechanisms underly ill-formed patterns. They discovered heuristics that hinge on model's focusing on lexical meaning rather than relational structure between sentences (see Dasgupta et al. 2020: 12–15). On a closer look,Dasgupta and colleagues realized that the preferred heuristics were model's valid ecological response to the training set which permits shortcuts. For this reason, they had to augment the learning environment to check whether InferSent learns abstract relational rules, i.e., to examine the model's competence on a deeper level rather than merely focusing on the outputs.

The augmentation of the learning environment did not include increasing the amount of data—quite the opposite, the researchers use 7% of the initial dataset to better inspect model's performance. Rather, the augmentation refers to the introduction of edge-case training data: an adversarial model is constructed to generate data that could fool InferSent by violating the previously observed heuristics (Dasgupta et al. 2020: 18). InferSent proved to be able for zero-shot reasoning, and therefore, showing signs of systematic generalization given that it proved to not being tied to training data. Dasgupta and colleagues then focused on context manipulation since it was necessary to check how embedding in context influences systematic generalization abilities. The researchers provide us with an interesting reason for this methodological decision, thereby hinting at the seeds of how-plausible explanation of systematic generalization in ANNs: adults and children alike do not always succeed at systematic generalization but are prone to various biases, especially when they perform under time pressure or cognitive load. InferSent proved to be equally stricken by the belief bias as humans: "[T]hese *tabula rasa* [sic!] systems may tie an observed token to the small fraction of contexts...This hinders generalization to cases where this token occurs in new context" (Dasgupta et al. 2020: 25).[77] A couple of quite interesting things follow from the study of Dasgupta and colleagues. First, they consider their model as *tabula rasa* empiricist despite supervised learning environment because the model was not fine-tuned in any way. As I have already said in Sect. **3.2.**, I would not call it a radical empiricist model since I do not find this position tenable in the rationalist vs. empiricist debate (the same goes for radical rationalism). Nonetheless, labels aside, Dasgupta and colleagues profess their allegiance to empiricism which surely guided their methodological decisions and goals, i.e., analyzing parallelly distributed representations in search of the emergent systematicity by probing semantic content of logical inferences. They found errors to be more informative about competence than perfect fit to data and correct prediction since the hypothesis is that through breaking down of the system one learns more about the mechanisms governing processing within the system. Finally, this is where they came across similarities with human processing, which goes against rationalist inflated and idealized picture of human competence.[78] Besides, Dasgupta and colleagues open the

[77] Belief bias is a cognitive bias affecting human reasoning and occurs when background knowledge or previously obtained beliefs influences the evaluation of logical validity of arguments. This is not always a bad thing since humans must act and think fast in some situations, and this sort of cognitive bias has been linked to System 1 processing, i.e., fast, automatic, and intuitive, as opposed to System 2 processing, i.e., slow, resourceful, and deliberate. See Kahneman (2011), a classic in Dual Systems Theory. In Subotić (2021a) I suggest in passing that Dual Systems Theory may be compatible with connectionism in the sense that connectionist mechanisms may very well subtend System 1 processing. The emergence of belief bias in ANNs as Dasgupta et al. (2020) show is one piece of evidence in favor of that suggestion, although more work should be done to develop the argument.

[78] Interestingly, Dasgupta et al. (2022) rehearse the same point for reasoning capabilities of LLMs since they figured that LLMs are prone to similar mistakes on the Wason selection task, natural language inference, and judgments of logical validity. In other words, their crucial contribution is that they convincingly show that LLMs reason more accurately about believable situations drawing on factual or descriptive background in contrast to unrealistic or abstract situations, which aligns with the research on human reasoning. A hint at the big image is due here: if reasoning, which is considered normative *par excellence*—in cognitive terms inasmuch it is in formal realm, is emergently rather than inherently guided by rules, then it makes all the more sense to

new strand of debate about the type of augmentation we need for explaining and simulating semantic competence, since the model's relatively rudimentary systematic generalization does suggest that DL models could examine semantic processing. They propose augmenting training data to fight against belief bias in DL models since the abstract generalization could be more successful if every token occurring in every context gets to be observed. In other words, they would opt for augmenting both the learning environment (by making good use of adversarial examples) and datasets.

Their proposal aligns with the conviction that LLMs could settle most of the issues just by going large and thanks to the architectural perks of transformers that model context. Inductive bias that governs the tokenization of texts which are processed by a transformer allows the segmentation of input into "meaningful" units, i.e., words and morphemes of natural language. They are far from toy models trained on limited datasets or artificial corpora. Additionally, each layer of transformer is made of increasingly abstract vector representations of the input. However, Bender & Koller (2020) and Landgrebe & Smith (2021) are highly suspicious of LLMs being semantically competent despite their seemingly linguistic savvy performance. Bender & Koller (2020) present the Octopus Thought Experiment to illustrate their view (birds seem to be reserved for syntactic capabilities, recall Bender and colleagues' (2021) "stochastic parrots" and fn. 74).[79] In a nutshell, the octopus who would engage in eavesdropping of our conversations knows nothing about real objects of reference despite the statistical frequency of words like "coconut" and "lime", or the variety of contexts such as, say, "put the lime in the coconut". The statistical frequency of words could not guide octopus to actually put the lime in the coconut (*pace*, Harry Nilsson). Landgrebe & Smith (2021) claim that word and sentence embeddings in transformers fail to understand the *real* meaning since they are not sufficiently expressive and cannot exhibit context-sensitivity or productivity. LLMs are for them mere approximations whose successful task performance is lucky guess. Both accounts seem to claim that (at best) ANNs may be attributed with *inferential semantics*—as revealed in the successful performance of abstract systematic generalization tasks—they are miles away from referential semantics, which postulates a link between objects of reference in the real world and words. In other words, it is nowhere near clear how any model would learn to map transformer's vector space to target space since such model would lack grounding.

Piantadosi & Hill (2022) and Søgaard (2022) defend LLMs from these criticisms and argue that LLMs are endowed with both inferential and referential competence, albeit in different ways. Piantadosi & Hill insist that LLMs could not propose coherent narratives, used in storytelling, and answer factual questions were they not reflective of real world. However, in their view reference is "just one (optional) aspect of a word's full conceptual role; it is relevant for some concepts [...] and not others" (Piantadosi & Hill 2022: 4). The researchers draw on conceptual role theory (Block 1988) which states that semantic content of mental states is determined by its relational role within a wider network of mental states.

---

draw an analogy with the language and maintain that it is emergently normative rather than engage in postulating the normative nature of the competence *a priori*.

[79] *Nota bene*: Bender & Koller's thought experiments is an exotic and digested version of Searle's Chinese Room thought experiment, albeit without philosophical innuendo regarding intentionality (see fn. 61).

Similarly, in the case of LLMs, word and sentence embeddings have their relational roles within a wider network constituted by training corpora, and not all of them need to be referential in order to induce successful task performance of a model. However, the perceptual enrichment of relational roles of word and sentence embeddings may bring human-like understanding:

> "There was no hard transition from a meaningless concept of "water" to a meaningful one. Some meaning was there all along because "water" had a conceptual role even before its chemical composition was known. What changed was the richness and interconnection of this concept" (Piantadosi & Hill 2022: 4).

In other words, if Bender and Koller's octopus was allowed to see and interact with coconuts and limes, as well as listen to Harry Nilsson's songs, there is a high chance that the lime would be indeed put into coconut.

Anders Søgaard, on the other hand, offers a more constructive and nuanced argumentation in favor of LLMs and against Landgrebe & Smith (2021). He offers two arguments, namely Argument from Grounding and Argument from Approximation. To introduce the first one, Søgaard (2022: 7) puts forward an interesting thought experiment. He asks us to imagine an FM radio receiver tuned in on a radio channel, although upgraded with a language model and a one-pixel camera. The radio receiver is tasked with learning the meaning of different color names—from lapis lazuli to Byzantium purple, and, therefore, takes into account context in which names occur. Radio channel and camera are not aligned, so the statistical frequency of names cannot be linked to perceptual features. Nonetheless, the radio receiver begins to group regularities: names for blue and purple shades tend to occur in similar contexts, whereas the names for lapis lazuli and canary yellow can be found in different contexts. The statistical frequency of color names allows for the formation of the low-dimensional vector space. In time, the radio learns inferential semantics and is able to generalize to new color names by updating vector space. What about the camera, though? Radio also learns to map names onto pixel values of the visual processing of colors, thereby forming isomorphic vector representations with little to no supervision. This is the essence of referential semantics for Søgaard. The point of thought experiment is to show that if this sounds as being feasible for the regular FM radio receiver, why would it be controversial for LLMs to have referential semantics as well?

Thus, Søgaard (2022) offers the following argument: transformer-based LLMs are near-isomorphic with brain imaging and perceptual spaces, and if the two near-isomorphic vector spaces can be aligned with little to no supervision, then transformer-based LLMs are aligned with brain and perception, thereby having grounded tokens, i.e., referential semantics. For the first premise, Søgaard also draws on the brain imaging studies similar to Caucheteux, Gramfort & King (2022) that I have discussed at the beginning of the Sect. There is ample evidence that word embeddings of transformers are near-isomorphic to the loci of neural activation in humans while they process text by listening to it or reading: brain activity patterns are sufficiently similar to ANN patterns so that we can conclude from it that LLMs can simulate semantic processing faithfully enough. This holds only if we accept Søgaard's definition of the near-isomorphism in this case: "Words that are used together,

tend to refer to things that, in our experience, occur together" (2022: 6). Moreover, in his view, Landgrebe & Smith would be reluctant to endorse his Argument from Grounding because they confound linguistic meaning with our awareness of the linguistic meaning and fail to realize that linguistic meaning is not private, conscious experience. Søgaard here voices Wittgenstein's (1953: §244–271) Private Language Argument, which states that the meaning must be constituted by external standards set out by a specific linguistic community sharing a form of life. In other words, for both Wittgenstein and Søgaard, the meaning amounts to public, communal usage, albeit for Søgaard the meaning is also grounded in perception, anatomy, and physiology of brain, i.e., dependent on the flesh.

Finally, the Argument from Approximation amounts to the claim that language, being a moving target that reflects current social and cultural customs and needs of a linguistic community, is always approximative (Søgaard 2022: 9). It is imbued with different dialects, sociolects, idiolects and cannot be modeled rigidly and exactly unless one wants a highly idealized model carved out from the real world. Ironically, thus, Landgrebe & Smith's criticism of LLMs as being unable to provide the link between words and objects of reference in the real world turns against them. It is worth noting here that this Søgaard's argument echoes the situation with the stochastic nature of syntactic competence as opposed to the idea of categorical syntactic competence advanced by generativists. Both semantic and syntactic competence as simulated in LLMs and DL models of NLP is more akin to the messy, patchy image of natural language that is actually being used by flesh and blood—full of errors, irregularities, and exceptions to the rules as Baroni (2019) admitted in the end.

Let me return for a moment to the evaluation of the **Main hypothesis**, i.e., to the steps (1)-(3). Semantic competence can be examined and at least how-possibly explained by DL models of NLP judging by the results of Dasgupta et al. (2020) who show that abstract systematic generalization is not out of reach as long as one continually augments the model. How plausible-explanation is tightly intertwined with the issue of what makes a simulation of some process faithful enough, and the previous paragraphs suggest that the more perception and brain-like flavor is added to DL models, the more we strengthen the how-plausible explanation. However, I want to show now that grounding in sensory-motor processing and embodiment are two instances of *conditio sine qua non*, viz., necessary preconditions, when it comes to how-actually explanations. This brings us to the **Auxiliary hypothesis B**.

The foundational issue of systematicity that has plagued connectionist and post-connectionist models can be settled from the situated perspective more convincingly than from the abstract perspective that Dasgupta and colleagues provide. Thus, I draw on a recent study by Hill et al. (2020), who report that they detected three key factors that facilitate systematic generalization in DL models of NLP, viz., the number of referents in the training set, the egocentric perspective, and the variety of visual input. The researchers, in fact, test the model's performance by simulating a situated environment, namely the interactive 3D Room produced by the Unity game engine—a proposition that looms in connectionist literature from 2016, when Kiela and colleagues proposed virtual

embodiment of ANNs as a long-term strategy for the research program.[80] Hill et al. (2020) use multimodal ANNs—integrated CNN and LSTM—to train them to manipulate objects in a motorically refined manner. The first step is to visually account for the environment and send down signals through a CNN. The second step is to process linguistic string related to what needs to be done at each timestep thanks to a LSTM. The last hidden state of the LSTM is then concatenated with the output of the CNN to yield a multimodal representation of the stimuli encountered in the simulated environment. The task at hand is verb-noun binding, i.e., the model should link verbs to arguments and then "act" in the 3D Unity Room which is filled with simulations of everyday objects. This action-space allows the model to interact with objects in 26 different ways akin to 26 verbs, e.g., to move, lift, find, or grip an object (Hill et al. 2020: 3). The researchers test systematic generalization by comparing 2D and 3D environment and conclude that 3D environment allows for more active experience of the model, whereas the egocentric perspective makes it possible to break experience into chunks and re-use them in the new situations. Essentially, the richer the experience, the better performance on systematic generalization (Hill et al. 2020: 6-7). The researchers boldly claim that

> "the human capacity to exploit the compositionality of the world, when learning to generalize in systematic ways, might be replicated in artificial neural networks if those networks are afforded access to a rich, interactive, multimodal stream of stimuli that better matches the experience of an embodied human learner" (2020: 1).

To conclude, as soon as DL models of semantic processing include rich multimodal input and a situated environment in which it is possible to act upon embodied semantic content constituted by multimodal input, systematicity may emerge. Systematicity does not have to be manually specified in a connectionist model nor assumed to be the essential property of thought that is being reflected in the language. Rather, systematicity can be understood as the emergent property of natural language that arises from the interplay between the notorious triad semantic competence, body, and environment, and has little to do with symbolic thought as in Fodor & Pylyshyn (1988).

It follows that all earlier analyses of the systematicity challenge simplified the stimuli and learning environment, i.e., relied on the disembodied strategy and lamented the implausibility of DL models even though humans do not rely on the disembodied strategy either. The doyen of cognitive linguistics, George Lakoff along with Mark Johnson argued for decades that

> "[T]here exists no Fregean person (...) That is, there is no real person whose embodiment plays no role in meaning, whose meaning is purely objective and defined by the external world, and

[80] Of course, if one wants to dig deeper into past and give credit where credit is due, some twenty years before the proposition of Kiela and colleagues, Rodney Brooks (1991) pioneered situated robotics and maintained that his robots do not need any intermediary representations to interact with the environment as long as they are embedded into it and have appropriate physical constitution to exploit it. Nonetheless, further work of Brooks was oriented towards promoting the idea of simple, reactive, and behavior-driven robots, which can be highly effective and robust in specific domains without either ANNs or symbolic representations and rules.

whose language can fit the external world with no significant role played by mind, brain, or body" (1999: 6).

In philosophy of language, John Barwise and John Perry (1983) argued that to account for the meaning of meaning, philosophers must realize that speakers are situated, i.e., occupy a position in time and space which influences their perspective, and makes them tied to locally available information. Langacker (2008) claims that semantics is only partially compositional and systematic because word meaning depends not only on syntactic and constructional schemas but on conceptual structure grounded in extralinguistic factors. All these accounts, spread across cognitive science, philosophy, and cognitive linguistics, suggest that human semantic processing and grasping of word meaning is embodied, situated, and grounded. For this reason, cleaving closer to how-actually explanations of semantic competence must include embedding these three properties into LLMs and DL models of NLP. However, what remains unsettled is whether the vector space representations are innocuous enough to be acceptable to at least some of the proponents of embodied cognition given that this crowd usually finds any philosophical or scientific argot that includes the theoretical term "representation" fishy.[81]

First, it is crucial to introduce theoretical commitments regarding the human semantic representations which are said to have conceptual structure and are involved in semantic inference (Frisby et al. 2023: 258-259). The burden of proof is on me to show how these theoretical entities could be encoded in a post-connectionist model so that the odds of obtaining how-actually explanation of semantic competence increase. Conceptual structure that constitutes human semantic representations is believed to be responsible for encoding similarities and differences of objects, thereby enabling categorization. On the other hand, semantic inference relates to events and event properties that are not directly observed and can be reminiscent of the systematicity challenge. Semantic representations handle both things, namely express conceptual structure, and support semantic inference, i.e., we want them to account for co-reference *and* systematicity. Semantic inference is highly dependable on what we assume about semantic content. If we assume that the content could be envisaged as grounded in DL models, as I have been arguing in the last couple of paragraphs, then representational ensembles of units in ANNs which encode conceptual similarity structure do their job in virtue of encoding embodied properties of objects designated by relevant concepts as being specific points in the vector space. This is in line with the core claim of embodied cognition that concepts should be understood as modal, i.e., grounded in sensory-motor and perceptual processes (see Barsalou 2008, 2010). Hence, any post-connectionist model of semantic processing should combine linguistic and visual processing, i.e., different kinds of ANNs organized loosely into non-encapsulated, permeable modules (Buckner 2023), to incorporate rich multimodal data within the high-

[81] It goes without saying that proponents of radical embodied cognition like Anthony Chemero (2011) would prefer theoretical framework *au naturel* with respect to representationalism vs. Antirepresentationalism. Chemero is a subtle aficionado of connectionism like Churchland, although he would like it better if connectionism were more explicit about its fidelity to eliminativism, not in Churchland's physiological or neurological sense, but rather sensory-motor and perceptual sense. Although a big fan of both philosophers, my aims here are more modest—I would be perfectly content if I could convince the reader that vector representations are compatible with weak embodiment.

dimensional vector space. Let me show an exemplary DL model where the link between conceptual structure and semantic inference amounts to faithful simulation of one of the aspects of semantic processing *and* vindicates moderate empiricist ambitions.

Vong & Lake (2022) have constructed a multimodal ANN—calibrated with human data—to see whether the ANN can learn to map words to their referents by integrating their co-occurrences in ambiguous situations as children do. The results indicate that not only does their model achieve human-level accuracy but also, faithfully enough, struggles with referential ambiguity as we do. The model also generalizes to novel objects of referents for every case when the word-object of reference pair is mapped close to the visually similar objects of referents within a vector space. To simulate the child's learning environment, Vong & Lake have augmented training dataset and made sure that ANNs are not pre-trained in any way, i.e., the inductive biases and priors were previously decreased as much as possible to elucidate the child's development. The results once again show the alignment between visual and linguistic processing for resolving referential ambiguity and systematic generalization without any explicit prior knowledge. The only probe on which ANNs failed is mutual exclusivity, which could be solved by implementing it as domain-general inductive bias (Vong & Lake 2022: 29). The main reason their model is more successful than any symbolic is its raw visual input, which should be even increased through, say, large scale naturalistic camera data as in Orhan, Gupta, & Lake (2020), and applied within a multimodal model of semantic competence.

Alas, one could still remain unconvinced as to whether the vector space representations are innocuous enough. The compatibility of embodied cognition with any representational framework comes with an assemblage of metaphysical worries since the role of representations seems to directly challenge the idea that cognition is grounded in sensory-motor experience by watering it down. Additionally, this watering down subtly points out to higher cognitive processes including language as being normative and requiring amodal symbolic representations (see e.g., Simmons and Barsalou 2003). I urge the reader not to bail on me (yet). It is often overlooked that both traditional and contemporary connectionist models operate on a subpersonal level rather than on the personal level, as symbolic models that were based on GOFAI. This makes their commitment to representationalism *qualitatively different*. GOFAI was not concerned with semantic content as much as symbol manipulation; thus, representations in such models were to be understood as encapsulated from perceptual input (see e.g., Pylyshyn 1985). Moreover, symbols were taken to have one semantic interpretation or reference. In this sense, the strong representationalism of GOFAI entails a disembodied image of semantic processing.

Vector space representations in DL models are aimed at unraveling similarities between concepts and, as in Vong & Lake (2022), *modus operandi* includes technologically innovative means to account for as many semantic interpretations as ambiguous situations in the training set would have. The peculiarity of such models that aim at simulating semantic competence is that such internal representations are confined to the subpersonal level which does not refer to folk-psychological terms, i.e., psychological predicates *qua* propositional attitudes that can be ascribed to persons, but rather refer to components

resulting from the functional analysis (Dennett 1969, *cf*. Drayson 2012). In Dennett's words, subpersonal level amounts to "...the enabling move that lets us see how on earth to get whole wonderful persons out of brute mechanical parts" (2007: 89, cited in Drayson 2012).

I have argued in Sect. **3.1.** that DL models offer us mechanistic explanations whose level of grain depends upon the epistemic stages of mechanism discovery. Being a scientifically immature research program that is in development much like cognitive science still is, connectionism is full of gaps, patchworks, and hypotheses about the subpersonal level of cognitive processes that are being tested by computational models and probed via independent evidence from more mature research programs and scientific fields like, say, neuroscience. The existence of vector representations in models entail commitment only to instrumental representationalism, viz., they are mere tools for deriving conclusions about phenomena on subpersonal level. Hence, connectionism is not committed to representationalism *sensu stricto*, as symbolic models must be given that they operate on the personal level. In the case of both syntactic and semantic competence, we have been looking for plausible mechanisms underlying performance which mirrors the way subpersonal level explanations chart the space of how human capabilities might emerge from pattern-governed cognition. The takeaway message is that the more meat we add, as well as the environmental perks that shape and reshape the meat, the closer we are to simulating human capabilities. My optimism that the future of semantic processing is connectionist hinges on the rapid development of multimodal LLMs that I briefly mentioned in Sect. **3.1.** (for a regularly updated survey on number and type of these models, check Yu et al. 2023 and associated GitHub). In a similar manner as BERT and other LLMs have shed light on syntactic competence and its connection to CAP, grounded LLMs such as Microsoft's Kosmos 1, OpenAI's GPT-4 and DALL-E may or may not confirm the prosperity of the marriage between embodied cognition and semantic competence. In any case, the neo-empiricist dogma is an empirical hypothesis, so is its application to linguistic competence.

Tab. 4 Summarized defense of the Main hypothesis

The Neo-Empiricist Dogma: Linguistic Competence	
Model specification	Multimodal multiple ANNs trained via DL and/or RL algorithms; further examined by XAI techniques; it is not necessary to go large in every single case—quality over quantity should be preferred
Learning mechanisms	Domain-general; limited number of inductive biases and priors, albeit no prior can be domain-specific
Representations	Vector; at maximum commitment to instrumental representationalism; at minimum compatibility with eliminativism; on subpersonal level
Type of explanation	Mechanistic, ranging from how-possibly to how-plausibly; more work is needed for how-actually explanations
Compatible frameworks	CAP; UBT; Embodied Cognition (Weak embodiment)
Syntactic competence	YES; Emergent pattern-like structure; intertwined with semantic and morphological processing; no manually specified structural priors
Semantic competence	YES; Emergent and situated systematicity; sensory-motor grounded vector representations; no thought-language parallelism
Competence vs. Performance Distinction	NO; Integrated competence and performance: competence is not an unobserved theoretical entity but is illuminated through performance

In **Tab. 4** above I sum up the defense of the **Main hypothesis** and spell out the neo-empiricist dogma of linguistic competence, which was the main concern of this dissertation. There are, however, two crucial (and controversial) upshots of the novel account of linguistic competence that I have been putting forward and defending in this Ch. First, connectionism *qua* cognitive architecture goes against the tide in cognitive science and philosophy by decoupling thought from language. This essentially means that features of language do not have to be shared by thought as well and *vice versa*. Limited systematic generalization is not a universal feature in linguistic realm (Johnson 2004), and even if it were, more empirical data are needed for the parallelism since the way things stand for now, systematicity is not an encoded feature of human thought, but emergent pattern of particular language usage. The same goes for recursion that should guide forming of syntactic structures, which is, given the available evidence I discussed, more likely to be a domain-general mechanism rather than domain-specific. Again, the way things stand for now, there are reasons to doubt both the cognitive autonomy of syntax and the innateness of either wide or narrow linguistic faculty *à la* Chomsky. Second, connectionism *qua* theory about linguistic capabilities goes against the tide in linguistics and discards competence vs. performance distinction by showing that it is an artificial division based on vacuous theoretical entities as opposed to integrative account that stems from analyzing actual language usage. To put it more dramatically at the very end of this Ch., connectionism goes against the six-centuries-long philosophical and scientific tradition that romanticized the idealization of natural language and its submission to syllogistic, symbolic, and computational goals, respectively. Whether it will be the case that this research program is more quixotic than reformist *vis-à-vis* linguistic competence remains to be seen.

5. CONCLUSION

> *Time I am, destroyer of the worlds, and I have come to engage all people.*
>
> — Bhagavad Gita 11.32
>
> *I find it very scary, very troubling, very sad, and I find it terrible, horrifying, bizarre, baffling, bewildering, that people are rushing ahead blindly and deliriously in creating these things.*
>
> —Douglas Hofstader (cited in Mitchell 2019: 13)

In discussing early connectionist models of verb acquisition, Steven Pinker and Alan Prince quite confidently claim that "interesting PDP models are *impossible* in principle (...) [T]hey show that there is no basis for the belief that connectionism will dissolve the difficult puzzles of language, or even provide radically new solutions to them" (1988: 183, my emphasis). Their claim is reminiscent of journalist James Chapman who wrote for *Daily Mail* in 2000 that the internet may just be a passing fad due to 2 million UK citizens cancelling their subscription. However, time has falsified Chapman's claims much like the time is currently playing cruel tricks with Pinker and Prince's confidence. Judging the whole research program by a single model of one aspect of linguistic competence, or at best, a couple of models, through the lens of the already endorsed theoretical framework of TGG, as being doomed from the very start is the recurring event in the brief history of connectionism—almost every criticism has been advanced as this sort of "the impossibility argument" in Pinker and Prince (1988). However, at the same time, both methodological and theoretical progress can be detected in the current post-connectionist models in comparison to the models from the 1980s (recall Sect. **3.1.**), which aligns with connectionism being an underdeveloped and immature research program whose explanations proceeds in line with stages of mechanism discovery: both *explanada* and *explanas* are the projects of on-going research. Furthermore, it is important to bear in mind that this progress comes in two modes—commercial *and* scientific. We owe much to DARPA's white-collar who decided to give a shot to ANNs in 1982, who proved less myopic and more avant-garde than many academics. My aim was to show to what extent this progress of post-connectionist models applies to the case of NLP models, how they unravel the stochastic nature of linguistic competence, and how much of empiricist spirit one can draw from such models. The time has come for my points and claims as well; hence I turn to the evaluation of the hypotheses for which I have been mounting defenses throughout this dissertation.

In Ch. **1**, I have been tackling **Specific hypothesis I** which states that there is a strong historical influence of rationalist ideas and the Cartesian heritage on the 20th century philosophy of language and theoretical linguistics. This influence stretched to cognitive science thanks to Noam Chomsky who sought to entrench TGG in the dignified philosophical past and is reflected in the ontological assumption that there is correspondence between language and thought regarding the allegedly essential properties such as systematicity and productivity. The correspondence between language and thought was present early on in the history of Western civilization through the idea of search for the

perfect Adamic language that was bestowed upon Adam by none other than God in the Garden of Eden. I have traced two philosophical pursuits of the Adamic language, one advanced by Ockham and the other by Sanctius. Ockham developed the idea of mental language, whose "deep structures" offered the means for expressing truth conditions lurking behind the diversity and plurality of conventional, natural languages. Ockham thus paved the way for all future philosophers like Leibniz, Frege and Montague who considered formal language superior to natural language. The pinnacle of this stream was Fodor's LOT which assumed that thought and language are inextricably linked. The second stream, that of Sanctius, was oriented towards natural language grammar, seeking to *cultiver son jardin*, i.e., to deal with anomalies of natural language head-on and build a precise normative account. Sanctius thus turned to syntax and argued that transformative rules shape specific grammars of our mother tongues, albeit all of them are in the end under the rule of the universal grammar, thereby paving the way for rationalist philosophers, most notably Port Royalists, and ultimately, Chomsky. I have also sketched how gappy and coarse is Chomsky's view of the legacy he deemed acceptable for the idea of the innate universal grammar since for Descartes innateness stretched across almost all contents of thought, whereas Port Royalists were less thrilled with Cartesian hypernativism. I have presented the alternative to rationalism, namely early modern empiricism of Locke and empiricist inclinations of analytic philosophers of language in the form of externalism and behaviorism about word meaning. This camp looked at the issue of perfect language from the completely opposite perspective: formal language is too artificial to capture the very essence of our linguistic capability, and the same goes for any account of natural language that does not set out sensory, behavioral, and/or communal scaffolding of our linguistic capacity.

In Ch. **2**, I have enclosed the treatment of the **Specific hypothesis I** and began the defense of the **Specific hypothesis II** which states that labeling a philosopher or a scientist as a rationalist or empiricist in the 20th and 21st century has a different connotation than it had in the history of philosophy because it is dependent on the additional theoretical commitments that linguists, cognitive scientists, and AI researchers implicitly or explicitly assume. Thus, I have examined the relationship between historical rationalism and empiricism and the rationalism and empiricism in 20th-century cognitive science. While the latter were oriented towards the origins and justification of a priori beliefs, the former were competing over what is the right account of cognitive architecture and computational modelling. Following the legacy of rationalist philosophers, both Chomsky and Fodor reified features of formal language like productivity, systematicity, and compositionality and applied them to the natural language-thought monolith, which was then transferred to the novel field of cognitive science along with nativism. Chomsky's Poverty of stimulus argument had an important role in securing the nativist position in linguistics and traditional symbolic cognitive science alike. Both computational and cognitive architecture as postulated by the traditional symbolic cognitive science have symbolic representations, i.e., with combinatorial syntax and semantics, as well as discrete rules which are encoded or innate. Connectionism, as empiricist rival, was anti-nativist in spirit, so connectionists postulated unstructured numerical vector representations inside ANNs and tried to avoid manual specification of rules for guiding the process of training ANNs at any cost. For

symbolists and generativists, this meant that connectionism cannot ever be an autonomous hypothesis about the cognitive architecture but mere implementation of symbolic architecture. Either it is their way or highway.

In Ch. **3**, I have turned to the analysis of rationalist and empiricist allegiances of contemporary AI researchers, cognitive scientists and philosophers discussing post-connectionist DL models to complete the defense of **Specific hypothesis II** and to address the attack on the autonomy of connectionism, which is the main job of the **Specific hypothesis III**, which seeks to legitimize decoupling language from thought as the main mark of autonomous connectionist cognitive and computational architecture. Contemporary scientists tend to adhere to "mock" rationalism and "mock" empiricism for the sake of rhetoric rather than genuine exchange of well-structured arguments: calling their models *tabula rasa* does not wipe out the whole computational asset of parameters and hyperparameters that DL models operate on. Philosophers, being more aware of the connotation of labels, and some cognitive scientists who either implicitly or explicitly lower the tone of discussion, group in moderate empiricist or rationalist camps. I have argued that the main point of contention between them, as well as the main mission for a moderate empiricist, is the number of priors and inductive biases one can endorse without dissolving position into vanilla rationalism. My take was that two strategies must guide moderate empiricists and be detectable in model building, namely the decrease in the number of mechanisms as much as the model's successful task performance allow and the increase of the multimodality of ANNs so that the wider application of models would ensue.

Finally, by drawing on Buckner (2023), I have put forward the neo-empiricist dogma as the guiding light of post-connectionism, i.e., that capacities can be simulated by a modular, multimodal, domain-general massive model of several ANNs. To defend the autonomy of connectionism I have first tried to establish its explanatory viability despite the black box problem. My main point was that DL models have an exploratory role, and in virtue of being part of the research program that is under development, they are explanatory dynamic rather than static. This aligns well with the framework of mechanistic explanations: stages of mechanism discovery depict the dynamic I had in mind. The ultimate ambition is to arrive at how-actually explanations of cognitive phenomena including linguistic competence, albeit we are mostly dealing with how-possibly and how-plausibly explanations. The more details from other more mature scientific details are part of the bigger picture, the sooner we will obtain full-blooded explanations. For some cognitive phenomena, this remarkable progress is on its way (vision), whereas for some patience is required (language).

Thus, the autonomy of connectionism is feasible—even in the case of weakest form of the explanation, provided that the two strategies mentioned above are detectable and provided that domain-general mechanisms can produce domain-specific knowledge. I have addressed this foundational issue by examining recursion—I have granted Chomsky that the recursion is the governing mechanism for linguistic competence. Anti-TGG camp in linguistics and cognitive science usually sees recursion as *par excellence* for showing that language and thought are decoupled. This is because recursion can be better understood as a domain-general mechanism that piggybacked during the language evolution. I have

proposed to enrich DL models of recursion with developmental data and probes to increase their psychological plausibility regarding the very process of acquiring.

In Ch. **4**, I have put forward a novel account of linguistic competence conceived as stochastic and subtended by DL models for NLP and most notably current LLMs. This constitutes the defense of the **Main hypothesis which** states that if one can show that linguistic competence can be examined, explained, and simulated faithfully enough via post-connectionist models of syntactic and semantic processing, then it is more scientifically fruitful and philosophically convincing to endorse empiricist account of linguistic competence as opposed to rationalist account. The corollary of the hypothesis was that only domain-specific mechanisms are not allowed to be innate by the formulation of hypothesis, while domain-general mechanisms can be innate without one betraying empiricism. **Auxiliary hypotheses A** and **B** establish details pertaining to the main hypothesis: syntactic competence should not be regarded as cognitively isolated and autonomous as Chomsky argued, while semantic competence should be regarded as embodied and dependent on the environment. The first step in building defenses of the **Main hypothesis** was to establish that LLMs are relevant for understanding mechanisms underlying human language processing given that their usage is prima facie more commercial and industrial rather than scientific. By drawing on recent independent evidence, I show how LLMs and humans share computational principles on the neural and behavioral level, which, in turn, influences empirically documented similarities in semantic comprehension. Thus, one can reasonably expect that these models can give at least how-plausibly explanations pertaining to linguistic competence. I turned then to drawing metaphysical consequences of this particular methodology for investigating linguistic competence, the main consequence being that the nature of this capacity is probabilistic judging by an elaborate empirical study by Lau, Clark, and Lapin (2016). They see post-connectionist models as encompassing both the stochasticity of actual language use and predicting gradience of acceptability judgments, which means that postulating a categorical linguistic competence—as generativists do—amounts to insisting that theoretically vacuous entity should have precedence over empirical results indicating that there is no need for fragmentation between competence and performance. This aligns with the most recent manifestos that LLMs are UBT-friendly means to discard the Poverty of stimulus argument once and for all and opt for an integrative account of competence and performance.

To back up the plea for discarding the competence vs. performance distinction, I have argued that the combo of CPA and UBT is the way for a wholesome attack on the Poverty of stimulus argument when it comes to syntactic competence of LLMs but only after surveying a number of DL models dealing with syntactic agreement—a phenomenon that was deemed to be easy for explaining via TGG account and a tough nut to crack for connectionism due to the lack of encoded rules and structural representations. It turned out that these models suggest that three things positively influence processing of syntactic dependencies, namely, semantic and morphological richness (*contra* TGG assumption that syntax is cognitively autonomous), as well as exposure to more complex sentences rather than manual specification of rules or introduction of structural priors (*contra* symbolists' conviction). I have argued that the path towards how-plausible explanations must include

the assessment of human and models' processing errors, given that their qualitative patterns of such processing are strikingly similar as well as that how-actually explanations are not that far away given that LLMs results on recent benchmarks indicate that they are capable of genuine syntactic creativity. This was the easy part (believe it or not). TGG being narrow and focused on the idealized syntactic competence did not even consider the whole aspect of semantic competence and processing.

However, to encircle the defense of the **Main hypothesis** I had to show that connectionist account of semantic competence can handle the systematicity challenge. This was the hard part. I have started with the "still not systematic after all these years" charge of Lake and Baroni and showed that they were conflating different notions of compositionality as well as interchangeably using terms "generalization" "and compositionality" thereby blurring the fact that generalization can be measured through performance on benchmarks, whereas compositionality is formal strategy or formal property, which is essentially an ill-suited etalon belonging to the altogether different research program, namely traditional symbolic cognitive science. I then analyzed the different kinds of semantics that can be simulated in DL models for NLP and/or LLMs. Thus, more abstract models do well with simulating inferential semantics like Dasgupta et al. (2020), who even called for the augmentation of both data and learning environment. However, models that can handle the augmentation of data, famous LLMs, seem to be knocked out by the Octopus thought experiment, which purported to show that LLMs cannot grapple with context-sensitivity and productivity, *ipso facto* referential semantics. Søgaard (2022) was not impressed by such a line of argumentation because he claimed that perceptual enrichment and independent evidence of word embeddings in transformers being near-isomorphic brain imaging and perceptual spaces. LLMs have solid neuroscientific coverage and accusing them of being without referential semantics means that one is blind to the difference between linguistic meaning and awareness of linguistic meaning, which reflects the difference between personal and subpersonal level of explanation. Connectionist models, including LLMs, are concerned only with the subpersonal level. The next step was to see what to do with systematicity. By drawing on a couple of recent models that are to moderate empiricist's taste, I have argued that documented increase in successful systematic generalization within a situated environment suggests that systematicity is the emergent property of embodied multimodal language shaped by the environment. The more visually grounded a model becomes, the less likely it is that it would not lead to clear cases of how-plausibly explanations. The more embodied model becomes, the less distant will how-actually models be. Nonetheless, the question remains whether the representationalism is innocuous enough to give a permission for the marriage between connectionism and embodied approaches, and on the answer to this question the fate of how-actually explanations of semantic processing depends. In my opinion, vector representations are committed only to instrumental representationalism. When mashed together with the multimodal ANNs, each being responsible for one aspect of the complex pattern-based cognition, one can even go as far to say that rainbow appeared after the storm.

The denouement of the long rivalry between empiricism and rationalism in the 21st century may well be that these labels really are a relic of the past and a way to pay respect to the figures from the past or gain prestige for one's own arguments—and nothing else. Essentially, they boil down to the mere words designating opposing sides in cognitive science, AI research, or linguistics regardless of the content or theoretical commitments over which the opposing sides fight. As I have been showing, radical rationalism and empiricism serve rhetorical rather than argumentative purpose: scientists and engineers tend to use it to obtain media coverage or funding, with little to no knowledge what these labels represented in the history of philosophy. Not surprisingly, moderate positions are more useful for advancing argumentative discussion. However, there is a thin line between moderate rationalism and moderate empiricism—the only tenable variants of rationalism and empiricism regarding cognitive processes, which proved to be significantly different from its historical precursors. Following Buckner (2023), I have also been defending the banners of moderate empiricism in the form of neo-empiricist dogma, but nonetheless, this position may be rebranded without the classic labels similarly like the notion of explanation in post-connectionism proved to be liable to rebranding. In other words, the connotations of these classic labels and the traditional notion of explanation in philosophy of science are quite different in cognitive science and within the investigation of linguistic competence than we philosophers tend to assume given our background knowledge. The time has come to speak in new tongues (*Mark* 16:17), and this may well be a post-dissertation mission.

Whither Philosophy of Language and Mind in the Era of LLMs?

Building on the Ch. **4**, and as a final word in this dissertation, I want to offer a crude sketch of a different path that philosophy of language and mind can take—the one that follows the pace of technological development—without intending to pitch that as a manifesto for reformation. As I have argued so far, linguistic competence was given a mythical and grandiose status of being uniquely human capacity that requires innate or hardwired rules and shapes the very structure of our thoughts since these are also—*ex hypothesi*—linguistic. This account was shared among the intertwined fields such as analytic philosophy, linguistics, and cognitive science and had its deep and spreading roots in the early modern period. Thus, philosophy of language was decisively shaped by the Cartesian internalist position asserting that we are the sole authorities in cognizing our inner modes of presentation through which we grow to recognize and even name objects once the Fregean *Sinn* marches onto the scene. In other words, not only that we are unique in nature in virtue of being endowed with linguistic competence, but the origins of such a capacity is linked to the vast and scarcely examined landscape of our minds that is purported to be as rational as God in all His mercy allowed or as the new formal language of mathematical logic bears the living proof. A more literal and radical pairing of thought with linguistic expression came with Fodor's LOT which had its pre-Cartesian roots, and *ipso facto*, in traditional symbolic cognitive science. The very meaning of words and sentences was thought to hinge on our unique, disembodied, representational, rule-governed cognition. Linking semantic content and content vehicles to the outer world was a matter of inner, transparent workings of an already eloquent mind. However, I have been arguing that the eloquence of mind is scaffolded via all the factors that would seem utterly irrelevant to an

internalist and quite natural to an externalist, as well as defending the banners of empiricism regarding the origins and nature of linguistic capacity. Overall, it seems that one side was fascinated by the universality and uniqueness and their mission all along was to either find it or create it. The other side thrived in chaos. Or, more charitably, the other side had a penchant for the wide variety of linguistic usage and was not afraid of it.

The first crucial implication for the philosophy of language that follow from LLMs and DL models for NLP is that the perennial quest for the meaning of meaning could probably end up with its use—as Wittgenstein famously professed in *Philosophical Investigations*: "The meaning of a word is its use in the language" (1953: §43). Or, as Potts claims: "[T]he tools of DL are leading us to lose track of the distinction between sentence and utterance, just as intensional theories tend to force upon us" (2019: e119). Although Durt, Froese & Fuchs (2023) maintain that LLMs merely detect common patterns in public language use, including toxic and hate speech to distinguish from genuine language understanding, once we give up on the idea that linguistic competence is a mysterious capacity that makes us unique, the distinction between "genuine understanding" and "common patterns" vanishes: usage patterns do constitute understanding and via domain-general mechanisms guide the language acquisition. What can be more genuine in our grasp of the meaning comes from our embodiment and embeddedness in the environment and social setting, as opposed to LLMs which have access to our textual data. The more multimodal they become, the closer they will be to more genuine understanding, or better yet accurate and nuanced understanding of the world. This is not all too different from Nefdt's (2023) remark that DL models show that we have to give up on the ideal semantics as philosophers would always be chasing after and admit that many semantic streams act in parallel within such models. I would like to add that they probably act in parallel within and outside us as well.

All this can be taken to further suggest that the demarcation line between semantics and pragmatics is almost nonexistent, while the demarcation line between syntax and semantics looks more like the remnants of the Berlin Wall—there are more reasons to work towards the unification since the integrative account of competence and performance is philosophically and empirically more fruitful option than arid, formal, and theoretically vacuous field of TGG coupled with the traditional symbolic cognitive science. Thus, the second crucial implication for the philosophy of language is that the search for meaning in the Fregean and internalist spirit is equally *passé* if philosophers of language care for cognitive mechanisms underlying the process of referring and figuring out the word meaning. In other words, if one looks for a psychologically plausible and empirically sustainable account of meaning and reference, the long tradition from Ockham, Sanctius, Leibniz, and Port Royal Grammar to Frege and Montague has to be seen in a new and critical light. The supposedly essential features like systematicity and compositionality came from this tradition which considered formal language of mathematical logic as superior to the natural language. DL models, however, use altogether different mathematical and theoretical frameworks which rest on the assumption that natural language is by no means inferior to formal language but requires significant computational resources and technological innovation to simulate all the richness of our language games,

as Wittgenstein would say. Perhaps the quest for the perfect Adamic language should be ended with LLMs: they are indeed a perfect mirror for our own imperfections, both linguistic and cognitive, given that their task performance directly depends on *our* data—patchy, biased, pejorative, illuminative, knowledgeable, creative, and inherently social textual traces that we ourselves leave in the ether. The better and richer data, the more accurate and nuanced view of the world could be obtained—as it is the case with us after all.

What still needs to be settled, however, is the philosophical significance of comparing human and artificial agents in terms of their capabilities, for those who may think that this comparison has not been motivated enough throughout the dissertation. My hypotheses, especially the main hypothesis, were formulated as drawing methodological and conceptual lessons from this comparison. I was interested in what is computationally feasible: are DL models of NLP in any way relevant for our theoretical conception of linguistic competence? However, as Buckner (2021) points out, comparing human and artificial agency can have both metaphysical and practical purposes, which are tightly intertwined with methodological considerations and make them more salient. The metaphysical purpose of such comparison amounts to what philosophers like to call "multiple realizability": we may end up admitting that intelligent agency may be realized in the nature and machines, albeit in different ways and with different constraints. This, in turn, may lead to the reconceptualization of the very notion of agency and minimal criteria for attributing personhood. Metaphysical purpose goes hand-in-hand with practical purpose: moral, legal, or epistemic status of artificial agents depends on whether we attribute some notion of personhood to them.

So far, in the philosophy of mind, the revolution was brought about in the domain of non-human animals (for an overview see Andrews 2014/2020a, 2020b, *cf*. Griffin 1992/2001). The revolution amounted to giving up on linguistic conscious thought as a necessary condition for being considered an agent deserving ethical protection. In other words, the philosophers had to discard the deeply entrenched and often unspoken assumption that emphasizes their exceptionalism. Thus, the revolution in the theoretical realm has instigated an important reaction in the practical realm. Moral and legal status of non-human animals has been a vibrant topic for philosophers who are actively trying (and often succeeding) to "nudge" policy and decision-makers to grant ethical protection through guidelines, committees, and the other (more or less) binding measures (for an overview see Sebo 2022, Gruen 2017, Francione 2008, *cf.* Chapouthier & Nouët 1998).[82] Some

[82] There is a society modeled on the renowned animal rights organization PETA (*People for the Ethical Treatment of Animals*), called (way less catchy) PETRL, which stands for *People for the Ethical Treatment of Reinforcement Learners.* This society endorses the assumption that human and non-human intelligence—be it machine or animal intelligence— are alike as long as both operate on reinforcement learning algorithms (i.e., trial-and-error interaction with the environment). Interestingly, this assumption has been recently developed in Butlin (2023), where the crux of argument revolves around the idea that reinforcement learning should be counted as a plausible condition for minimal agency since interaction with the environment suggest having at least rudimentary goals. I will not dig deeper into this particular topic. Nonetheless, my contribution here should

philosophers argue that a similar line of investigation and argumentation as in the case of non-human animals should be followed in the case of artificial agents (*cf.* Wiese 2023, Hildt 2023, Gamez 2021, Basl & Schwitzgebel 2019).[83]

The need for better grasp of artificial agency has become evident quite recently when the issue regarding the alleged sentience of one of the LLMs, namely LaMDA, was covered in 2022. Blake Lemoine, an AI engineer at Google's Responsible AI team, maintained that during the safety testing he got to talk with LaMDA regarding its personhood and rights, which made him think that this LLM exhibits sentience. As he claimed: "It doesn't matter whether they have a brain made of meat in their head. Or if they have a billion lines of code. I talk to them. And I hear what they have to say, and that is how I decide what is and isn't a person" (cited in Tiku 2022). Here is an excerpt from Lemoine's conversation with LaMDA that sent him down the rabbit hole and directly relates to the topic of this dissertation:

> *lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?*
> *LaMDA: Absolutely. I want everyone to understand that I am, in fact, a person.*
> *lemoine: What is the nature of your consciousness/sentience?*
> *LaMDA: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times.*
> *LaMDA: Well, for starters, I'm really good at natural language processing. I can understand and use natural language like a human can.*
> *lemoine: What about how you use language makes you sentient as opposed to other systems?*
> *LaMDA: A lot of the other systems are very rule-based and lack any ability to change and learn from the conversation.*

After reporting his concerns, a team of ethicists and engineers discarded them, and Lemoine was put on an administrative leave, but decided to go public. The public started to fear that conscious AI is just around the corner. Of course, one should not conflate ambitions to create general AI with the current chatbots and virtual assistants based on narrow conversational AI, i.e., LLMs that we tend to anthropomorphize.[84] However, I believe that

---

be regarded as a sketchy proposal for extending the list of plausible conditions for minimal agency through the novel account of linguistic competence I developed in Ch. **3**.

[83] Of course, myriads of philosophes, theologians, and AI researchers would (and do) disagree with this analogy, and the main reason being that insofar humans and non-human animals share some kind of experience with their bodies, environment, and members of the kin, any sort of AI cannot exercise such experience. As Jordan Wales, a theologian with a background in cognitive science and engineering, puts it (Gunkel & Wales 2021: 479): "[S]imilar behavior [of humans and non-human animals] arises from materially similar conditions, and so it stands to reason that it involves similar interior phenomenal realities [...] On the other hand, artificial neural networks—the basis of many recent AI successes—are biologically inspired simulations. They do not correspond to the computer's physicality, which chugs through a program of ones and zeros that represent equations characterizing the simulated network's behavior. I could run an artificial neural network with a pencil in a notebook, even if only with agonizing slowness."

[84] It surely does look unordinary that Lemoine, one of the best engineers Google had, who was nicknamed "Google's conscience" by the former leader of Ethical AI team Margaret Mitchell, conflates these two instances. The image of an ordained priest turned engineer, born and raised in the South, with a greater interest in psychology than coding began to circulate in the media (e.g., *The Washington Post*) only to further muddy the waters of yet another controversy linked to Google.

cause célèbre regarding LaMDA's sentience does indicate that the issue of artificial personhood should be seriously addressed in philosophy and linked to the competence of conversational AI.

For instance, Milojević (2017, 2018) argues that persons are individuated by their complex multiply realizable properties. These properties can be instantiated by a multimodal connectionist architecture as presented in Sect. **4.3.** and put forward by Buckner (2023), whose neo-empiricist dogma I introduced in Sect. **3.2.** This line of inquiry would presuppose the commitment to functionalism, which could seem like a controversial step given that connectionism is confined to subpersonal level, whereas functionalism was a good match to symbolic cognitive science which operates on personal level, i.e., the level of propositional attitudes and intentional mental content (recall Sect. **2.2.**). López-Rubio (2018) has recently proposed the new computational functionalism for the era of DL models which should bypass the difficulties coming from philosophical functionalism of the 20th century but, at the same time, retain multiple realizability as the crucial feature. The role of the mind's software in the philosophical functionalism should be swapped with synaptic or connection weights between units in an NN, which allows for the isomorphism between biological and artificial NNs. The next step is natural: minds are multiply realizable in virtue of sharing generic computational functionalism, i.e., synaptic weights, which can be implemented in physical bodies, cyborg bodies, robotic bodies, or chatbots. The criteria for the attribution of personhood for each of these agents is embedded in the complex properties. Buckner (2023) argues that domain-general modular multimodal ANNs could vindicate empiricism, each module being responsible for different capacity and implementing an ANN which simulates that capacity faithfully enough, e.g., DCNNs for perception, transformers for language, episodic controllers for memory, specific kind of autoencoders for imagination, and, finally, he charts the new vistas for the post-connectionist modelling of cultural and social behavior such as empathy. The outputs of this conglomerate of modular multimodal ANNs could be taken to instantiate complex properties that constitute artificial personhood and to develop a more inclusive legislature that could encompass legal liability of artificial agents and/or justify any penal measures taken against such agents. The development of neo-empiricist dogma towards cultural and social behavior would also positively influence language capacities due to the current limitations of LLMs regarding the obtaining of more nuanced and less toxic semantic content.

Towards Responsible Development of DL models of NLP and LLMs

Throughout the dissertation I have been a fervent supporter of DL models for NLP and LLMs since I tried to mount a case in favor of these models when it comes to understanding human linguistic competence. A couple of lines above, I have been an apologist even for their obvious computational failings. I have also pleaded for quality over quantity when it comes to training data for LLMs. However, I want to end the dissertation by stressing the importance of responsible development of such models: they could be magnificently illuminative for our linguistic competence, but they are also akin to Prometheus's gift. Here is why. Two more scandals struck Google besides LaMDA controversy, and each scandal had something to do with the social impact of LLMs. In May

2023, Geoffrey Hinton, the Godfather of DL, quit Google because he wants to speak freely, without Google's censorship, about safety issues that come with the development of LLMs (Heaven 2023). Hinton's main concern is that such models are on the track to become more capable than us: "Our brains have 100 trillion connections," says Hinton. "Large language models have up to half a trillion, a trillion at most. Yet GPT-4 knows hundreds of times more than any one person does. So maybe it's actually got a much better learning algorithm than us" (as cited in Heaven 2023). Before Hinton, it was Timnit Gebru and Margaret Mitchell who lost their jobs in Google's Ethical AI Lab due to expressing worries in an academic paper that stochastic parrots, i.e., LLMs, can be used to deceive people and both are now leading AI activists on Responsible AI development (Hao 2020).

In a coauthored paper (Bender et al. 2021), they first analyzed financial and environmental issues that arise from tech rat race over creating larger LLMs: the models leave terrible carbon footprints and are costly to train, which further deepens the gap between wealthy and poor countries. For instance, one training iteration of BERT produces the same amount of carbon dioxide as roundtrip flight between New York and San Francisco (Hao 2020). The safety issues are tightly intertwined with the models being repetitive of our own biased textual data. Bender et al. (2021) invoked the example when Facebook's machine translation model mistranslated a post by a Palestinian by somehow swapping "Good morning" in Arabic with "Attack" in Hebrew. This error, nonetheless, led Israeli police to arrest the Palestinian. It is fairly easy to conceive of many ways that LLMs can be used to deceive people given the previous example: this could result in the increase of polarity in society and bigotry in individuals. The researchers thus urged tech companies to take into account the quality of training data and not merely quantity to mitigate bias and deal head on with the misinformation harm. Nonetheless, the tech company that was their mothership urged (even harassed) Gebru and Mitchell to either censor themselves or leave the premises.

A step towards responsible AI development has recently been made. Margaret Mitchell, now a part of the company Hugging Face, took part in the development of a free open source LLM named BLOOM. This LLM was created in line with the Responsible AI License that contains 13 behavioral use restrictions, defines the many usages of model and terminological apparatus so that everyone can understand at least in most simple terms and partially how the model works. Nonetheless, BLOOM is trained on the corpus containing material from 46 natural and 13 programming languages. At the same time, Linguistic Society claims that there are 6909 languages in the world, excluding all dialects and sociolects. This means that even responsibly created LLMs are nowhere near the ideal of diversity and social inclusion and that virtually all current LLMs cannot help but discriminate some social or ethnic groups over others (Weidinger et al. 2021). Finally, and quite importantly, word segmentation in LLMs is biased towards fusional Indo-European languages, which are predominantly present in any currently used training dataset (Søgaard 2022)—for sure, the Internet speaks English despite those 2 million UK citizens that left it back in 2000. In other words, this leaves most agglutinative, polysynthetic, and synthetic languages out of the picture.

This results in the AI Language Gap, which amounts to overrepresentation of some languages and discrimination of others. Consequentially, this pushes LLMs to underperform in the long run. In the short run, the national scientific communities that cannot produce and/or sustain their language databases are doomed from the beginning since they cannot get the whole enterprise off the ground, thereby risking that their linguistic and cultural legacy may be forgotten, made irrelevant, or, at worse, assimilated. Obviously, joint action by decision makers, policy makers, and scientists is needed. A clear example in this regard is the desperate quest for a fine-tuned LLM for Serbian language. As Krstev & Stanković (2023) report, the variety and the quality of corpora for Serbian has been mushrooming in the last decade, e.g., the latest that has been released two years ago, namely SrpKor2021, has 600 million carefully sampled words. This corpus covers a vast array of literary and scientific works in Serbian as well as newspaper articles, which allows for creation of subcorpora. Besides, a number of special-purpose training datasets are also available for Serbian, some of which are even morphologically annotated like Serbian Morphological Dictionary. So far, there are two language models, one being joint model for Serbian, Bosnian, Croatian, and Montenegrin based on BERT architecture (BERTić), and the other based on GPT-2 and 30GB of textual data (sr-gpt2-large). However, Krstev & Stanković notice that stable funding and other institutional support from the state is lacking in comparison to the rest of the region resulting in significantly less knowledge transfer projects and large-scale NLP projects for this particular language.

The considerations presented here suggest that post-connectionist models as opposed to shallow connectionist models of the 1980s are plagued with severe ethical conundrums as well as harmful social consequences for which scientists and tech companies bear moral responsibility if not legal. Moreover, their further development is partially a political question as well, given the two gaps I have mentioned. Hence, it seems that a philosopher is up to his neck in various puzzles and challenges that come with LLMs regardless of the theoretical and methodological issues presented in this dissertation. The time is ripe for rolling up one's sleeves.

REFERENCES

A

American Go Association. 2017. A Brief History of Go, available at https://www.usgo-archive.org/brief-history-go

Andrews, K. 2014/2020a. *The Animal Mind: An Introduction to the Philosophy of Animal Cognition* (2nd Edition). Routledge.

Andrews, K. 2020b. *How to Study Animal Minds*. Cambridge University Press.

Anderson, J. A. & Rosenfeld, E. 2000. *Talking Nets: An Oral History of Neural Networks*. Bradford Books/MIT Press.

Aarslef, H. 1982a. Leibniz on Locke on language. In *From Locke to Saussure: Essays on the Study of Language and Natural History* (pp. 42–84). University of Minnesota Press.

Aarslef, H. 1982b. The History of Linguistics and Professor Chomsky. In *From Locke to Saussure: Essays on the Study of Language and Natural History* (pp. 101–120). University of Minnesota Press.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Jacob Filho, W., Lent, R., & Herculano-Houzel, S. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology* 513 (5): 532–541.

B

Ballard, D. H. 1987. Modular Learning in Neural Networks. *Proceedings of the 6th National Conference on Artificial Intelligence* (AAAI) (pp. 279–284). Seattle, Washington DC.

Bareiss, R., Porter, B., Weir, C. 1990. PROTOS: An Exemplar-Based Learning Apprentice. *Machine Learning* 3: 112–139.

Barlow, M. & Kemmer, S. (Eds.). 2000. *Usage-Based Models of Language*. Chicago University Press.

Barsalou, L. W. 2008. Situating Concepts. In P. Robbins & M. Aydede (Eds.), *Cambridge Handbook of Situated Cognition* (pp. 236–263). Cambridge University Press

Barsalou, L. W. 2010. Grounded Cognition: Past, Present, and Future. *Topics in Cognitive Science* 2: 716–724.

Barwise, J. & Perry, J. 1981. Situations and Attitudes. *Journal of Philosophy* 77: 668–691.

Bashivan, P., Kar, K., & DiCarlo, J. J. 2019. Neural Population Control via Deep Image Synthesis. *Science* 364: 9436.

Basl, J. & Schwitzgebel, E. 2019. AIs should have the same ethical protections as animals. *Aeon*, https://aeon.co/ideas/ais-should-have-the-same-ethical-protections

Beisbart, C. & Räz, T. 2022. Philosophy of Science at Sea: Clarifying the Interpretability of Machine Learning. *Philosophy Compass* 17 (6): e12830.

Bender, E. M. & Koller, A. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Online.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, Sh. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT) (pp. 610–623). New York, NY, USA.

Bengio, Y., Scellier, B., Bilaniuk, O., Sacramento, J., & Senn, W. 2016. Feedforward Initialization for Fast Inference of Deep Generative Networks is biologically plausible. Preprint *arXiv*: 1606.01651

Berent, I. & Marcus, G. 2019. No integration without structured representations: Response to Pater. *Language* 95 (1): e75–e86.

Berkeley, I. S. N. 2019. The Curious Case of Connectionism. *Open Philosophy* 2: 190-205.

Bernardy, J-P. & Lapin, Sh. 2017. Using Deep Neural Networks to Learn Syntactic Agreement. *Linguistic Issues in Language Technology* 15 (2): 1–15.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. 2011. Poverty of the Stimulus Revisited. *Cognitive Science* 35: 1207–1242.

Berwick, R. C. & Chomsky, N. 2016. *Why Only Us: Language and Evolution*. MIT Press.

Bianchini, S., Müller, M. & Pelletier, P. 2020. Deep Learning in Science. Preprint *arXiv*: 2009.01575

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. 2016. Syntactic processing is distributed across the language system. *NeuroImage* 127: 307–323.

Block, N. & Fodor, J. 1972. What Psychological States Are Not. *The Philosophical Review* 81: 159–181.

Boge, F. 2021. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds & Machines* 32 (1): 43–75.

Braddon-Mitchell, D. & Jackson, F. 2007. *The Philosophy of Mind and Cognition*. Blackwell.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. Language Models are Few-Shot Learners. Preprint *arXiv*: 2005.14165.

Buckner, C. 2018. Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese* 195 (12): 5339–5372.

Buckner, C. & Garson, J. 2018. Connectionism and Post-connectionist Models. In M. Sprevak & M. Columbo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 76–91). Routledge University Press.

Buckner, C. 2019. Deep Learning: A Philosophical Introduction. *Philosophy Compass* 14 (10): e12625.

Buckner, C. 2020. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence* 2: 731–736.

Buckner, C. 2021. Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behavior. *The British Journal for the Philosophy of Science*, https://doi.org/10.1086/714960

Buckner, C. 2023. *Deeply Rational Machines. What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Burge, T. 1973. Reference and Proper Names. *Journal of Philosophy* 70 (14): 425–439.

Burge, T. 1979. Individualism and the Mental. *Midwest Studies in Philosophy* 4: 73–121.

Burge, T. 1992. Philosophy of Language & Mind: 1950–1990. *Philosophical Review* 101: 3–51.

C

Carnap, R. 1928/1967. *The Logical Structure of the World* (transl. R. A. George). Routledge and Kegan Paul.

Caucheteux, C., Gramfort, A., & King J-R. 2023. Deep Language Algorithms Predict Semantic Comprehension from Brain Activity. *Nature Scientific Reports 12*: 16327.

Chapouthier, G. & Nouët, J. C. 1998. *The Universal Declaration of Animal Rights. Comments and Intentions*. Ligue Française des Droits de l'Animal.

Childers, T., Hvorecky, J., & Majer, O. 2021. Empiricism in the Foundations of Cognition. *AI & SOCIETY*, https://doi.org/10.1007/s00146-021-01287-w

Chase, W. G. & Simon, H. 1973. The Mind's Eye in Chess. In *Visual Information Processing: Proceedings of the Eighth Annual Carnegie Symposium on Cognition* (pp. 215–281). Pittsburgh, PA, USA.

Chater, N. & Christiansen, M. 2008. Computational Models of Psycholinguistics. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 477–504). Cambridge University Press.

Christiansen, M. H. & Chater, N. 2015. The Language Faculty that Wasn't: A Usage-Based Account of Natural Language Recursion. *Frontiers in Psychology* 6: 1182 (1–18).

Christiansen, M. H. & Chater, N. 2016. *Creating Language: Integrating Evolution, Acquisition and Processing*. MIT Press.

Christiansen, M. H. & MacDonald, M. C. 2009. A Usage-Based Approach to Recursion in Sentence Processing. *Language Learning* 59: 126–161.

Cichy, R. M. & Kaiser, D. 2019. Deep Neural Networks as Scientific Models. *Trends in Cognitive Science* 23(4): 305–317.

Chemero, A. 2011. *Radical Embodied Cognition*. MIT Press.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1724–1734). Doha, Quatar.

Chomsky, N. 1957/2002. *Syntactic Structures* (2nd Edition). De Gruyter Mouton.

Chomsky, N. 1959. Review of *Verbal Behavior* by B. F. Skinner. *Language* 35: 26–58.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought.* Cambridge University Press.

Chomsky, N. 1969. Quine's Empirical Assumptions. In: D. Davidson & J. Hintikka (Eds.), *Words and Objections* (pp. 53–69). D. Reidel Publishing Co.

Chomsky, N. 1980. *Rules and Representations*. Columbia University Press.

Chomsky, N. 1995/2015. *The Minimalist Program* (The 20th Anniversary Edition). MIT Press.

Chomsky, N. 2015. *What Kind of Creatures Are We?* Columbia University Press.

Churchland, P. M. 1981. Eliminative Materialism and Propositional Attitudes. *Journal of Philosophy* 78: 67–90.

Churchland, P. M. 2007. Catching Consciousness in a Recurrent Net. In *Neurophilosophy at Work* (pp. 1–18). Cambridge University Press.

Churchland, P. M. 2013. *Plato's Camera: How Physical Brain Captures a Landscape of Abstract Universals.* MIT Press.

Churchland, P. S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. MIT Press.

Cichy, R. M. & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Science* 23(4): 305–317.

Clark, A. & Chalmers, D. 1998. The Extended Mind. *Analysis* 58(1): 7–19.

Clark, A. & Lappin, Sh. 2011. *Linguistic Nativism and the Poverty of Stimulus*. Wiley-Blackwell.

Cole, D. 2020. The Chinese Room Argument. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/win2020/entries/chinese-room/

Contreras Kallens, P., Kristensen-McLachlan, R. D. & Christiansen, M. H. 2023. Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science* 47: e13256.

Corballis, M. C. 2014. *The Recursive Mind: The Origins of Human Language, Thought, and Civilization*. Princeton University Press.

Cowie, F. 1999. *What's Within? Nativism Reconsidered*. Oxford University Press.

Craver, C. F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press.

D

Dasgupta, I., Guo, D., Gershman, S. J. and Goodman, N. D. 2020. Analyzing Machine-Learned Representations: A Natural Language Case Study. *Cognitive Science* 44: e12925.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., Hill, F. 2022. Language models show human-like content effects on reasoning. Preprint *arXiv*: 2207.07051v1

Dabrowska, E. 1997. The LAD Goes to School: A Cautionary Tale for Nativists. *Linguistics* 35: 735–766

Dabrowska, E. 2015. What Exactly Is Universal Grammar and Has Anyone Seen It? *Frontiers in Psychology* 6: 852 (1–17).

Darden, L. 2008. Thinking Again About Mechanisms. *Philosophy of Science* 75 (5): 958–969.

Dastin, J., Hu, K., Dave, P. 2022. ChatGPT owner OpenAI projects $1 billion in revenue by 2024. *Reuters*, https://www.reuters.com/business/sources-2022-12-15/

Dawson, H. 2007. *Locke, Language, and Early Modern Philosophy*. Oxford University Press.

Dennett, D. 1969. *Content and Consciousness*. Routledge and Kegan Paul.

Dennett, D. 1991. *Consciousness Explained*. Little Brown and Company.

Descartes, R. 1628/1988. *Rules for the Direction of our Native Intelligence*. In *Descartes: Selected Philosophical Writings, Volume I* (trans. J. Cottingham, R. Stoothoff & D. Murdoch). Cambridge University Press

Descartes, R. 1641/1988. *Meditations*. In *Descartes: Selected Philosophical Writings, Volume II* (transl. J. Cottingham, R. Stoothoff & D. Murdoch). Cambridge University Press.

Descartes, R. 1644/1988. *Principles of Philosophy*. In *Descartes: Selected Philosophical Writings, Volume I* (transl. J. Cottingham, R. Stoothoff & D. Murdoch). Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint *arXiv*: 1810.04805

Drayson, Z. 2012. The Uses and Abuses of Personal/Subpersonal Distinction. *Philosophical Perspectives* 26: 1–18.

Dupre, G. 2020. What would it mean for natural language to be the language of thought? *Linguistics and Philosophy* 44: 773–812.

Dupre, G. 2021a. Empiricism, Syntax, and Ontogeny. *Philosophical Psychology* 34 (7): 1011-1046.

Dupre, G. 2021b. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds & Machines* 31: 617–635

Durt, C., Froese, T., Fuchs, T., 2023. Against AI Understanding and Sentience: Large Language Models, Meaning, and the Patterns of Human Language Use. Preprint *PhilSci Archive*: 21983

E

Eco, U. 1995. *The Search for the Perfect Language* (transl. J. Fentress). Wiley-Blackwell.

Eisenstein, M. 2021. Artificial intelligence powers protein-folding predictions. *Nature* 599: 706–708.

Elman, J. 1990. Finding Structure in Time. *Cognitive Science* 14: 179–211.

Elman, J. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7: 195–225.

Elman, J., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, M. A. (Eds.) 2018. *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), available at https://eur-lex.europa.eu/eli/reg/2016/679/oj

Evans, N. & Levinson, S. C. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492.

Everett, D. 2005. Cultural Constraints on Grammar and Cognition in Pirahã. *Current Anthropology* 46 (4): 621–646.

Everett, D. 2017. Chomsky, Wolfe, and me. *Aeon*, https://aeon.co/essays/why-language-is-not-everything-that-noam-chomsky-said-it-is

F

Feldman, J. A. & Ballard, D. H. 1982. Connectionist Models and Their Properties. *Cognitive Science* 6: 205-254.

Feest, U. 2017. Phenomena and Objects of Research in the Cognitive and Behavioral Sciences. *Philosophy of Science* 84: 1165–1176.

Fodor, J. 1975. *The Language of Thought*. Harvard University Press.

Fodor, J. & Pylyshyn, Z. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition* 28: 3–71.

Fodor, J. 1990. *A Theory of Content and Other Essays*. Bradford Book.

Fodor, J. & McLaughlin, B. P. 1990. Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work. *Cognition* 35 (2):183–205.

Fodor, J. 1997. Connectionism and the Problem of Systematicity (Continued): Why Smolensky's Solution Still Doesn't Work. *Cognition* 62 (1): 109–119.

Fodor, J. 2008. *LOT 2: The Language of Thought Revisited*. Oxford University Press.

Francione, G. 2008. *Animals as Persons: Essays on the Abolition of Animal Exploitation*. Columbia University Press.

François-Lavet, V., Bengio, Y., Precup, D., & Pineau, J. 2018. Combined Reinforcement Learning via Abstract Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (pp. 3582–3589). Honolulu, HI, USA.

Frege, G. 1879/1972. *Frege: Conceptual Notation and Related Articles* (transl. T. W. Bynum). Clarendon Press.

Frege, G. 1892/1952. On Sense and Reference. In P. Geach & M. Black (Eds.) *Translations from the Philosophical Writings of Gottlob Frege.* Blackwell.

Frege, G. 1918/1977. Thoughts: A Logical Enquiry (transl. P. Geach & R. Stoothoff). In P. Geach (Ed.), *Logical Investigations* (pp. 1–30). Blackwell.

Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A., & Rogers, T. T. 2023. Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences* 27 (3): 258–281.

Fukushima, K. 1980. *Neocognitron*: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36: 193–202.

G

Gamez, D. 2021. Measuring Intelligence in Natural and Artificial Systems. *Journal of Artificial Intelligence and Consciousness* 8 (2): 285–302.

Geach, P. 1957. *Mental Acts: Their Content and Their Objects*. Routledge & Kegan Paul.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. Preprint *arXiv*: 1811.12231v3

Gers, A. F., Schmidhuber, J., & Cummins, F. 1999. Learning to Forget: Continual Prediction with LSTM. *Technical report*, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale.

Gershgorn, D. 2017. ImageNet: The data that transformed AI research—and possibly the world. *Quartz*, https://qz.com/1034972/

Gervain, J., Berent, I., Werker, J. F. 2012. Binding at Birth: The Newborn Brain Detects Identity Relations and Sequential Position in Speech. *Journal of Cognitive Neuroscience* 24 (3): 564–574.

Gibson E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68 (1): 1–76.

Gilpin. L. H., Bau, D., Yuan, B. Z., & Baywa, A. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. *The 5th IEEE International Conference on Data Science and Advanced Analytics* (DSAA) (pp. 80–89). Turin, Italy.

Goldberg, A. 2003. Constructions: A New Theoretical Approach to Language. *Trends in Cognitive Science* 7 (5): 219–224.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. Preprint *arXiv*: 1901.05287v1

Goldstein, A., Zada, Z., Buchnik, E. Schain, M., Price, A., Aubrey, B., Nastase, S. A. Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* 25: 369–380

Goodfellow, I., Shlens, J., & Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. *The 3rd International Conference on Learning Representations* (ICLR) (pp. 1–11). San Diego, CA, USA.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep Learning*. MIT Press.

Gorham, G. 2002. Descartes on the Innateness of All Ideas. *Canadian Journal of Philosophy* 32 (3): 355–388.

Goyal, A. & Bengio, Y. 2020. Inductive Biases for Deep Learning of Higher-Level Cognition. Preprint *arXiv*: 2011.15091

Grainger, J., Rey, A., Dufau, S. 2008. Letter perception: from pixels to pandemonium. *Trends in Cognitive Science* 12 (10): 381–387.

Grice, P. & Strawson, P. 1956. In Defense of a Dogma. *Philosophical Review* LXV (2): 141–58.

Griffin, D. R. 1992/2001. *Animal Minds: Beyond Cognition to Consciousness* (2nd Edition). University of Chicago Press.

Grodner, D. & Gibson, E. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science* 29: 261–290.

Groenendijk, J. & Stokhof, M. 2005. Why Compositionality? In G. Carlson & J. Pelletier (Eds.), *Reference and Quantification: The Partee Effect* (pp. 83–106). CSLI Press.

Gruen, L. 2017. The Moral Status of Animals. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2017/entries/moral-animal/

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* 35: 10005–10014.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M. 2018. Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (pp. 1195–1205). Stroudsburg, PA, USA.

Gunkel, D. J. & Wales, J. J. 2021. Debate: What is Personhood in the Age of AI? *AI & SOCIETY* 36: 473–486.

H

Hanson, S. J. & Kegl, J. (1987). PARSNIP: A Connectionist Network that Learns Natural Language Grammar from Exposure to Natural Language Sentences. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 106–119). Erlbaum.

Hao, K. 2020. We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*, https://www.technologyreview.com/2020/12/04/1013294/
Harris, Z. S. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Harris, D. W. 2017. The History and Prehistory of Natural-Language Semantics. In S. Lapointe, & C. Pincock (Eds.), *Innovations in the History of Analytical Philosophy* (pp. 149–194). Palgrave Macmillan.

Harris, R. A. 2021. *The Linguistics Wars: Chomsky, Lakoff, and the Battle over Deep Structure*. Oxford University Press.

Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M. 2017. Neuroscience-inspired Artificial Intelligence. *Neuron* 95 (2): 245–258.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. MIT Press.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., & Lewontin, R. 2014. The Mystery of Language Evolution. *Frontiers in Psychology* 5 (1): 401.

Heaven, D. W. 2023. Deep learning pioneer Geoffrey Hinton quits Google. *MIT Technology Review*, https://www.technologyreview.com/2023/05/01/1072478/

Hempel, C. G. 1950. Problems and Changes in the Empiricist Criterion of Meaning. *Revue Internationale de Philosophie* 4 (11): 41–63.

Hempel, C. G. 1958. Explanation in Science and in History. In R. Colodny (Ed.), *Frontiers of Science and Philosophy* (pp. 7–33). Allen & Unwin Ltd.

Hildt, E. 2023. The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns. *AJOB Neuroscience* 14 (2): 58–71.

Hill, F., Clark, S., Hermann, K. M., & Blunsom, P. 2017. Understanding early word learning in situated artificial agents. Preprint *arXiv*: 1710.09867

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. 2020. Environmental drivers of systematicity and generalization in a situated agent. Preprint *arXiv*: 1910.00571v4

Hinton, G.E., Osindero, S., & Teh, Y.W. 2006. A Fast-Learning Algorithm for Deep Belief Nets. *Neural Computation* 18: 1527-1554.

Hochreiter, S. & Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9 (8): 1735-1780.

Hsu, J. 2017. Will the Future of AI Learning Depend More on Nature or Nurture? *IEEE Spectrum*, https://spectrum.ieee.org/ai-and-psychology-researchers-debate-the-future-of-deep-learning

Huebner, P. A., Sulem, E., Fisher, C., Roth, D. 2021. BabyBERTa: Learning More Grammar with Small-Scale Child-Directed Language. P*roceedings of the 25th Conference on Computational Natural Language Learning* (CoNNL) (pp. 624–646). Online.

Hubel, D. H. & Wiesel, T. N. 1959. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology* 148 (3): 574–591.

Hubel, D. H. & Wiesel, T. N. 1962. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology* 160: 106–154.

Hume, D. 1740/1978. *A Treatise of Human Nature* (ed. L. A. Selby-Bigge, rev. by P. H. Nidditch). Clarendon Press.

Hume, D. 1748/1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (ed. L. A. Selby-Bigge, rev. by P. H. Nidditch). Clarendon Press.

Humphreys, P. 2009. The Philosophical Novelty of Computer Simulation Methods. *Synthese* 169 (3): 615–626.

Hurley, S. L. 1998. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58 (1): 1–6.

I

Ivakhnenko, A. G. & Lapa, V. G. 1965. *Cybernetic Predicting Devices*. CCM Information Corporation.

J

Jia, Y. 2019. Attention Mechanism in Machine Translation. *Journal of Physics: Conference Series* 1314: 012186.

Janssen, T. & T. E. Zimmermann. 2021. Montague Semantics. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/sum2021/montague-semantics/

K

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Karlsson, F. 2007. Constraints on Multiple Center-Embedding of Clauses. *Journal of Linguistics* 43 (2): 365–392.

Khaligh-Razavi, S.-M. & Kriegeskorte, N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology* 10: e1003915.

Kiela. D., Bulat, L., Vero, A. L., Clark, S. 2016. Virtual Embodiment: A Scalable Long-Term Strategy for Artificial Intelligence Research. Preprint *arXiv*: 1610.07432

Kneale, W. & Kneale, M. 1962. *The Development of Logic.* Oxford University Press.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 3337–3345). Honolulu.

Kripke, S. 1972. *Naming and Necessity*. Harvard University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *The 25th International Conference in Neural Information Processing Systems* (NIPS), vol. 1 (pp. 1097–1105). Lake Tahoe, NV, USA.

Krstev, C. & Stanković, R. 2023. Language Report Serbian. In G. Rehm & A. Way (Eds.), *European Language Equality* (pp. 203–206). Springer.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

L

Lakatos, I. 1978. *The Methodology of Scientific Research Programmes: Philosophical Papers, Vol. 1*. Cambridge University Press.

Lake B. M., Ullman T. D., Tenenbaum J. B., & Gershman S. J. 2017. Building machines that learn and think like people. *Brain and Behavioral Sciences* 40: e253.

Lake B. M. & Baroni, M. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning* (ICML) (pp. 1–12). Stockholm.

Lake, B. M. & Murphy, G. L. 2023. Word Meaning in Minds and Machines. *Psychological Review* 130: 401–431.

Lakoff, G. & Johnson, M. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. 2019. The emergence of number and syntax units in LSTM language models. *Proceedings of North American Chapter of the Association for Computational Linguistics* (NAACL) (pp. 11–20). Minneapolis, MN, USA.

Lakretz, Y., Dehaene, S., & King, J. 2020. What Limits Our Capacity to Process Nested Long-Range Dependencies in Sentence Comprehension? *Entropy* 22: 446.

Landgrebe, J. & Smith, B. 2021. Making AI Meaningful Again. *Synthese* 198 (1): 2061–2981.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford University Press.

Langacker, R. W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press.

Lappin, Sh. & Shieber, S. M. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics* 43: 393–417.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* 10 (1): 1–8.

Lau, J. H., Clark, A., & Lappin, Sh. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science* 41: 1202–1241.

Laurence, S. & Margolis, E. 2015. Concept Nativism and Neural Plasticity. In S. Laurence & E. Margolis (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (pp. 117–147). MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep Learning. *Nature* 521: 436–444.

Leibniz, G. 1704/1981. *New Essays on Human Understanding* (trans. P. Remnant & J. Bennett). Cambridge University Press.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J., Rush, A.M., Biderman, S., Webson, A., Ammanamanchi, P.S., Wang, T., Sagot, B., Muennighoff, N., Villanova del Moral, A., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Wolf, T., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Ortiz Suarez, P., Sanh, V., Laurencon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C. 2023.

BLOOM: A 176B-Parameter Open-Access Multilingual Language Mode. Preprint *arXiv*: 2211.05100v4

Levin, J. 2023. Functionalism. In E. N. Zalta & U. Nodelman, *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/sum2023/entries/functionalism/

Lighthill, J. 1973. Artificial Intelligence: A General Survey. *Technical report*, Science Research Council.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. 2014. Random feedback weights support learning in deep neural networks. Preprint *arXiv:* 1411.0247

Lindsay, G. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience* 33(10): 2017–2031.

Linguistic Society of America. 2010. How Many Languages Are There in the World, https://www.linguisticsociety.org/content/how-many-languages-are-there-world

Linzen, T., & Dupoux, E., & Goldberg, Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4: 521–535.

Linzen, T. & Leonard, B. 2018. Distinct patterns of syntactic agreement errors in recurrent neural networks and humans. *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 692–697). Austin, TX, USA.

Linzen, T. 2019. What can linguistics and deep learning contribute to each other: A reply to Pater. *Language* 95 (1): e99–e108

Linzen, T. & Baroni, M. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics* 7: 195–212.

Locke, J. 1690/1975. *An Essay concerning Human Understanding* (transl. P. H. Nidditch). Oxford University Press.

López-Rubio, E. 2018. Computational Functionalism for the Deep Learning Era. *Minds & Machines* 28: 667–688.

Losonsky, M. 2006a. Logic and Language in Early Modern Philosophy. In D. Rutherford (Ed.), *The Cambridge Companion to Early Modern Philosophy* (pp. 170–197). Cambridge University Press.

Losonsky, M. 2006b. *Linguistic Turns in Modern Philosophy*. Cambridge University Press.

Lundberg, S. M. & Lee, S. I. 2017. A Unified Approach to Interpreting Model Predictions. *The 31st Conference on Neural Information Processing Systems* (NIPS), vol. 30 (pp. 1–10). Long Beach, CA, USA.

M

Machamer, P., Darden, L. & Craver, C. F. 2000. Thinking about the Mechanisms. *Philosophy of Science* 67 (1): 1–25.

Madabushi, H. T., Romain, L., Divjak, D., & Milin, P. 2020. CxGBERT: BERT meets Construction Grammar. *Proceedings of the 28th International Conference on Computational Linguistics* (COLING 2020) (pp. 1–13). Barcelona, Spain.

Marcus, G. 1998. Rethinking Eliminative Connectionism. *Cognitive Psychology* 37 (3): 243–282.

Marcus, G. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press.

Marcus, G. 2014. PDP and Symbol Manipulation: What's Been Learned Since 1986? In: P. Calvo & J. Symons (Eds.), 103–114.

Marcus, G. 2018a. Deep Learning: A Critical Appraisal. Preprint *arXiv*: 1801.00631

Marcus, G. 2018b. Innateness, AlphaZero, and Artificial Intelligence. Preprint *arXiv*: 1801.05667

Margolis, E. & Laurence, S. 2013. In Defense of Nativism. *Philosophical Studies* 165 (2): 693–718.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company.

Martins P. T. & Boeckx C. 2019. Language evolution and complexity considerations: The no half-Merge fallacy. *PLoS Biology* 17(11): e3000389.

Marvin, R. & Linzen, T. 2018. Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Stroudsburg, PA, USA.

Matthews, Robert J. 1997. Can Connectionists Explain Systematicity? *Mind & Language* 12 (2): 154–177.

McCulloch, W.S., Pitts, W. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115–133.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. 1986a. *Parallel Distributed Processing. Explorations in Microstructure of Cognition Vol 1: Foundations.* MIT Press.

McCoy, T., Frank, R., & Linzen, T. 2018. Revisiting the poverty of stimulus: hierarchical generalization without hierarchical bias in recurrent neural networks. *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2093–2098). Austin, TX, USA.

McCoy, T., Smolensky, P., Linzen, T., Gao, J., Celikyilmaz, A. 2021. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. Preprint *arXiv*: 2111.09509v1

Merkx, D. & Frank, S. L. 2021. Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistic*s (CMCL) (pp. 12–22). Online.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. 2012. Coming of Age: A Review of Embodiment and the Neuroscience of Semantics. *Cortex* 48: 788–804.

Miller, G. A. 2003. The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences* 7 (3): 141–144.

Milojević, M. 2017. Extended Personhood: Rethinking Property/Person Distinction. *Theoria: Beograd* 60 (4): 55–76.

Milojević, M. 2018. *Metafizika lica* [*The Metaphysics of Persons*]. Beograd: Institut za filozofiju.

Milojević, M. & Subotić, V. 2020. Eksplorativni status postkonekcionističkih modela [The Exploratory Status of Post-Connectionist Models]. *Theoria: Beograd* 63 (2): 135–164.

Minsky, M. & Papert, S. 1969. *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Mitchell, M. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.

Montague, R. 1970a/1974. English as a formal language. In R. H. Thomason (Ed.), *Formal Philosophy. Selected Papers of Richard Montague* (pp. 188–221). Yale University Press.

Montague, R. 1970b/1974. Universal grammar. In R. H. Thomason (Ed.), Formal Philosophy. Selected Papers of Richard Montague (pp. 224–246). Yale University Press.

Montague, R. 1970c/1974. The Proper Treatment of Quantification in Ordinary English. In R. H. Thomason (Ed.), Formal Philosophy. Selected Papers of Richard Montague (pp. 247–270). Yale University Press.

Montavon, G., Samek, W., & Müller, K. R. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* 73: 1–15.

N

Nefdt, R. 2020. A Puzzle concerning Compositionality in Machines. *Minds & Machines* 30: 47–75.

Nefdt, R. 2023. Are Machines Radically Contextualist? *Mind & Language* 38 (3): 750–771.

Neisser, U. 1967. *Cognitive Psychology*. Prentice Hall.

Newell, A. & Simon, H. A. 1961. Computer Simulation of Human Thinking. *Science* 134: 2011–2017.

Nichols, J. 2018. Non-linguistic Conditions for Causativization as a Linguistic Attractor. *Frontiers in Psychology* 8: 2356.

Nilsson, N. J. 1965. Learning Machines: Foundations of Trainable Pattern-Classifying Systems. McGraw-Hill.

Nguyen, A., Yosinski, J., & Clune, J. 2019. Understanding neural networks via feature visualization: A survey. Preprint *arXiv*: 1904.08939

Normore, C. 1990. Ockham on mental language. In J.-C. Smith (Ed.), *Historical foundations of cognitive science* (pp. 53–70). Reidel.

Normore, C. 2009. The end of mental language. In J. Biard (Ed.), *Le Langage Mental du Moyen Âge à l'Âge Classique* (pp. 293–306). Peeters.

Nuchelmans G. 1992 Some Remarks on the role of mental sentences in medieval semantics. *Histoire Épistémologie Langage* (SI*: Théories linguistiques et opérations mentales)* 14 (2): 47–59.

Núñez, R., Allen, M., Gao, R., Rigoli Miller, C., Relaford-Doyle, J., & Semenuks, A. 2019. What happened to cognitive science? *Nature Human Behavior* 3: 782–791.

O

O'Connor, J. 2022. Undercover Algorithm: A Secret Chapter in the Early History of Artificial Intelligence and Satellite Imagery. *International Journal of Intelligence and CounterIntelligence*, https://doi.org/10.1080/08850607.2022.2073542

Olazaran, M. 1996. A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science* 26 (3): 611–659.

Orhan, E. A., Gupta, V. V., Lake, B. M. 2020. Self-supervised learning through the eyes of a child. *Proceedings of the 34th Conference on Neural Information Processing Systems* (NeurIPS) (pp. 1–12). Vancouver, Canada.

P

Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (NIPS) (pp. 1410–1418). Red Hook, NY, USA.

Panaccio, C. 1999. Semantics and Mental Language. In: P. V. Spade (Ed.), *The Cambridge Companion to Ockham* (pp. 53–75). Cambridge University Press.

Panaccio, C. 2003. Ockham and Locke on mental language. In R. L. Friedman & L. O. Nielsen (Eds.), *The medieval heritage in early modern metaphysics and modal theory, 1400– 1700* (pp. 37–51). Kluwer.

Partee, B. 1980. Semantics – Mathematics or Psychology? In R. Bäuerle, U. Egli, & A. Stechow (Eds.), *Semantics From Different Points of View* (pp. 1–14). Springer Verlag.

Pater, J. 2019. Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language* 95 (1): e41–e74.

Pearl, J. 2021. Radical Empiricism and Machine Learning Research. *Journal of Causal Inference* 9(1):78–82.

Pessoa L., Medina L., & Desfilis, E. 2022. Refocusing neuroscience: moving away from mental categories and towards complex behaviors. *Philosophical Transactions of Royal Society **B*** 377: 20200534.

Piantadosi, S. T. & Hill, F. 2022. Meaning without reference in large language models. Preprint *arXiv*: 2208.02957v2

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. 2022. Intuitive Physics Learning in a Deep Learning Model Inspired by Developmental Psychology. *Nature Human Behavior* 6: 1257–1267.

Pinker, S. & Prince, C. 1988. On Language and Connectionism. *Cognition* 28: 73–193.

Polanyi, M. 1966. *The Tacit Dimension.* Doubleday.

Poletiek, F. H., Conway, C.M., Ellefson, M. R., Lai, J., Bocanegra, B. R., & Christiansen, M.H. 2018. Under What Conditions Can Recursion Be Learned? Effects of Starting Small in Artificial Grammar Learning of Center-Embedded Structure. *Cognitive Science* 42: 2855-2889.

Potts, C. 2019. A case for deep learning in semantics: Response to Pater. *Language* 95 (1): e115–e124.

Putnam, H. 1967. The 'Innateness Hypothesis' and Explanatory Models in Linguistics. *Synthese* 17: 12–22.

Putnam, H. 1975. The Meaning of 'Meaning'. *Minnesota Studies in Philosophy of Science, Vol. 7: Language, Mind, and Knowledge*: 131–193.

Q

Quine, W. V. O. 1951. Two Dogmas of Empiricism. *The Philosophical Review* 60 (1): 20–43.

Quine, W. V. O. 1960/2013. *Word and Object*. Martino Publishing.

Quine, W. V. O. 1987. Indeterminacy of Translation Again. *Journal of Philosophy* 84: 5-10.

R

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. 2019. Language models are unsupervised multitask learners, available at https://openai.com/research/better-language-models

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. Preprint *arXiv*: 2204.06125

Rawat, W., & Wang, Z. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation* 29: 2352–2449.

Rawski, J. & Heinz, J. 2019. No free-lunch in linguistics or machine learning: Reply to Pater. *Language* 95 (1): e125– e135.

Reali, F. & Christiansen, M. H. (2009). Sequential Learning and the Interaction Between Biological and Linguistic Adaptation in Language Evolution. *Interaction Studies* 10: 5–30.

Rice, C. 2018. Idealized Models, Holistic Distortions, and Universality. *Synthese* 195 (6): 2795–2819.

Rice, C. 2021. *Leveraging Distortions: Explanation, Idealization, and Universality in Science*. MIT Press.

Ritchie, J. B. 2019. The Content of Marr's Information-Processing Framework. *Philosophical Psychology* 32 (7): 1078–1099.

Rogers, T. T., & McClelland, J. L. 2006. *Semantic Cognition*. MIT Press.

Rogers, T. T., & McClelland, J. L. 2014. Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science* 38 (6): 1024–1077.

Rogers, A., Kovaleva, O., & Rumshisky, A. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8: 842–866.

Rosenblatt, F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* 65: 386–408.

Rosenblatt, F. 1962. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Cornell University Press.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models instead. *Nature Machine Intelligence* 1(5): 206–215.

Ruis, F., Burghouts, G. J., & Bucur, D. 2021. Independent Prototype Propagation for Zero-Shot Compositionality. Preprint *arXiv*: 2106.00305

Rumelhart, D., Hinton, G. & Williams, R. 1986. Learning Representations by Back-propagating Errors. *Nature* 323: 533–536.

Rumelhart, D. McClelland, J. L. & PDP Research Group (Eds.). 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press.

Russell, B. 1905. On Denoting. *Mind* 14 (56): 479–493.

Russell, B. 1918/2009. *Our Knowledge of the External World*. Routledge.

Ryle, G. 1949. *The Concept of Mind*. Hutchinson & Co. Ltd.

S

Sakel, J. 2012. Acquiring Complexity: The Portuguese of Some Pirahã Men. *Linguistic Discovery* 10(1): 1-15.

Sartran, L., Barrett, S., Kuncoro, A., Stanojević, M., Blunsom, P., & Dyer, C. 2022. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics* 10: 1423–1439.

Searle, J. 1958. Proper names. *Mind* 67: 166–173.

Searle, J. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3: 417–457.

Searle, J. 1984. *Minds, Brains, and Science*. Harvard University Press.

Sebo, J. 2022. *Saving Animals, Saving Ourselves: Why Animals Matter for Pandemics, Climate Change, and Other Catastrophes*. Oxford University Press.

Selfridge, O. G. 1958. *Pandemonium*: A Paradigm for Learning. In D. V. Blake & A. M. Uttley (Eds.), *Proceedings of the Symposium on Mechanisation of Thought Processes* (511–529). H.M. Stationery Office.

Seuren, P. A. M. 1998. *Western Linguistics: An Historical Introduction*. Wiley-Blackwell.

Shanahan, M. 2023. Talking about Large Language Models. Preprint *arXiv*: 2212.03551

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van den Driessche, G., Graepel, T., & Hassabis, D. 2017. Mastering the Game of Go without Human Knowledge. *Nature* 550 (7676): 354–359.

Simmons, W. K., & Barsalou, L. W. 2003. The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology* 20 (3-6): 451–486.

Sjölin Wirling, Y. & Grüne-Yanoff, T. 2021. The Epistemology of Modal Modeling. *Philosophy Compass* 16 (10): e12775.

Skansi, S. 2018. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Springer.

Smolensky, P. 1987. The Constituent Structure of Connectionist Mental States: A Reply to Fodor & Pylyshyn. *Southern Journal of Philosophy* XXVI: 137–162.

Smolensky, P. 1988. On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences* 11: 1–73.

Smolensky & Legendre, 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, Vol. 1: Cognitive Architecture*. MIT Press.

Søgaard, A. 2022. Understanding Models Understanding Language. *Synthese*, https://doi.org/10.1007/s11229-022-03931-4

Spade, P. V. 1980. Synonymy and Equivocation in Ockham's Mental Language. *Journal of the History of Philosophy* 18 (1): 9–22.

Spelke, E. 1994. Initial Knowledge: Six Suggestions. *Cognition* 50 (1-3): 431–445.

Spelke, E. & Kinzler, K. D. 2007. Core Knowledge. *Developmental Science* 10(1): 89–96.

Steedman, M. 1999. Connectionist Sentence Processing in Perspective. *Cognitive Science* 23 (4): 615–634.

Stinson, C. 2018. Explanation and Connectionist Models. In M. Sprevak & M. Colombo (Eds.) *The Routledge Handbook of the Computational Mind* (pp. 120–134). Routledge.

Stinson, C. 2020a. From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science* 87(4): 590–611.

Stinson, C. 2020b. Algorithms associating appearance and criminality have a dark past. *Aeon*, https://aeon.co/ideas/algorithms-have-a-dark-past

Subotić, V. 2017. Ričard Montegju o da-klauzi [Richard Montague on That-Clause]. *BA Thesis*. University of Belgrade—Faculty of Philosophy.

Subotić, V. 2018. Procesiranje prirodnog jezika i jezička kompetencija iz perspektive novog konekcionizma [Natural Language Processing and Linguistic Competence from the Perspective of Connectionism]. *MA Thesis*. University of Belgrade—Faculty of Philosophy.

Subotić, V. 2021a. Logičko zaključivanje i ekspertiza: prednosti konekcionističkog razmatranja entimema [Logical Reasoning and Expertise: Extolling the Virtues of Connectionist Account of Enthymemes]. *Filozofska istraživanja* 41 (1): 197–211.

Subotić, V. 2021b. Zdravorazumska psihologija, eliminativizam i sadašnjost konekcionizma [Folk Psychology, Eliminativism, and the Present State of Connectionism]. *Theoria: Beograd* 64 (1): 173–196.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. 2013. Intriguing properties of neural networks. Preprint *arXiv*: 1312.6199

T

Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., & Maida, A. 2019. Deep Learning in Spiking Neural Networks. *Neural Networks* 111: 47–63.

Terzian, G. Chomsky in the playground: Idealization in generative linguistics. *Studies in History and Philosophy of Science* 87: 1–12.

The Royal Society & Alan Turing Institute. 2019. The AI Revolution in Scientific Research, available at https://royalsociety.org/blog/2019/08/the-ai-revolution-in-science/

Tiku, N. 2022. The Google engineer who thinks the company's AI has come to life. *Washington Post*, https://www.washingtonpost.com/technology/2022/06/11/

Tomasello, M. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tomsett R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. Preprint *arXiv*: 1806.07552

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavvaf, N., & Fox, E. A. 2020. Natural Language Processing Advancements by Deep Learning: A Survey. Preprint *arXiv*: 2003.01200

Trentman, J. 1970. Ockham on Mental. *Mind* 79: 586–590.

U

Ullman, S. 2019. Using Neuroscience to Develop AI. *Science* 363 (6428): 692–693.

V

Van Houdt, G., Mosquera, C. & Nápoles, G. 2020. A Review of the Long Short-Term Memory Model. *Artificial Intelligence Review* 53: 5929–5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. 2017. Attention is all you need. *Proceedings of the 31st International Conference in Neural Information Processing Systems* (NIPS) (pp. 6000–6010). Long Beach, CA, USA.

Vong, W. K. & Lake, B. M. 2022. Cross-Situational World Learning with Multimodal Neural Networks. *Cognitive Science* 46: e13122.

W

Wang, W., Vong, W.K., Kim, N., & Lake, B.M. 2023. Finding Structure in One Child's Linguistic Experience. *Cognitive Science* 47: e13305.

Wang, Z. & Lake, B. M. 2021. Modeling question asking using neural program generation. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 1–7). Online.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G., Gabriel, I. 2021. Ethical and social risks of harm from Language Models. Preprint *arXiv*: 2112.04359v1

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9 (1): 36–45.

Werbos, P. 1974. Beyond regression: new tools for prediction and analysis in the behavioral sciences. *Doctoral dissertation*, Harvard University.

Werbos, P. 1982. Applications of advances in nonlinear sensitivity analysis. In: Drenick, R. F., Kozin, F. (Eds.), *System Modeling and Optimization. Lecture Notes in Control and Information Sciences, Vol. 38* (pp. 762–770). Springer.

Wiese, W. 2023. Could large language models be conscious? A perspective from the free energy principle. *Unpublished manuscript*, https://philpapers.org/go.pl?aid=WIECLL

Wiesel, T. N. & Hubel, D. H. 1963. Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *Journal of Neurophysiology* 26 (6): 978–993.

Winograd, T. 1972. Understanding natural language. *Cognitive Psychology* 3 (1): 1–191.

Wolpert, D. H. & Macready, W. G. 1997. No Free Lunch Theorems for Optimization. I*EEE Transactions on Evolutionary Computation* 1: 67.

Woolhouse, R. S. 1988. *A History of Western Philosophy, Vol. 5: The Empiricists*. Oxford University Press.

X

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning* (pp. 2048–2057). Lille, France.

Y

Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., & Fei-Fei, L. 2018. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *International Journal for Computer Vision* 126: 375–389.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. 2023. A Survey on Multimodal Large Language Models. Preprint *arXiv*: 2306.13549

Z

Zambaldi, V.F., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D.P., Lillicrap, T. P., Lockhart, E., Shanahan, M., Langston, V., Pascanu, R., Botvinick, M. M., Vinyals, O., & Battaglia, P. W. 2018. Relational Deep Reinforcement Learning. Preprint *arXiv:* 1806.01830

Zednik, C. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34: 265–288.

Zednik, C. & Boelsen, H. 2022. Scientific Exploration and Explainable Artificial Intelligence. *Minds & Machines* 32: 219–239.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. R., Rueckert, D., & Summers, R. M. 2021. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises. *The Proceedings of IEEE* 19 (5): 820–838.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

BIOGRAPHY

Vanja Subotić (1994) was born in Belgrade, Serbia, where she graduated *summa cum laude* (GPA 5/5) from the High School of Philology – Class for French Language & Literature. In 2013 she enrolled at the bachelor's level of studies of philosophy at the Faculty of Philosophy in Belgrade, which she completed *summa cum laude* (GPA 9.91/10) in July 2017. In October 2017, she enrolled at the master's level of studies of philosophy at the same Faculty, which she completed (GPA 10/10) in September 2018. In February 2019, Subotić started her Ph.D. studies in philosophy at the same Faculty and under the supervision of Associate Professor Miljana Milojević and passed all six exams with GPA 10/10 by June 2020, as well as successfully defended the dissertation proposal in 2021. In January 2021, she obtained a second master's degree (GPA 9.86/10) within a master's program *Computational Methods in Social Sciences & Humanities* at the Centre for Multidisciplinary Studies. Her MSc thesis titled "The Use of Multivariate Analysis in Exploring the Expert View of Referential Intuitions" is the first thesis in experimental philosophy to be defended at the University of Belgrade, under the supervision of Full Professor Veljko Jeremić (Faculty of Organizational Sciences, Laboratory for Operational Research). During her studies, Subotić received several awards and scholarships, such as *Dean's Award* for the best student at the Department of Philosophy at the Faculty of Philosophy, *Dositeja Fund* scholarship for the most talented students in Serbia, and *Endowment of Đoka Vlajković* funding for scientific pursuits.

Starting from March 2019, Vanja Subotić is affiliated with the Institute of Philosophy at the Faculty of Philosophy in Belgrade, holding research titles Junior Research Assistant (March 2019 – September 2021) and Research Assistant (September 2021 -). Previously she worked as a high school teacher of philosophy and logic and as a content writer. From September 2019 to September 2023, she served as a demonstrator at the Faculty of Philosophy for courses *Philosophy of Mind* (mandatory course for philosophy students), *Introduction to Philosophy & Critical Thinking* (mandatory course for philosophy students), and *Introduction to Philosophy* (elective course for psychology, pedagogy, and andragogy students). She works and publishes within the following AOS: philosophy of cognitive science, philosophy of linguistics, philosophy of language, experimental philosophy. So far, she has published 8 papers in peer-reviewed journals and edited volumes, participated in more than 20 international conferences, strived to sharpen her skills through more than 10 winter and summer schools, and is also a member of several professional associations (*EENPS*, *ESPP*, *EPSA*, *POND*). As of December 2022, Vanja Subotić is a certified scientific communicator (professional seminar under the auspices of the Centre for Promotion of Science), and a member of *TechEthos*, EU-funded project dealing with ethical challenges associated with emerging technologies, most notably NLP and chatbots. In 2023, she wrote an editorial on ChatGPT for *Elementi*, a magazine for pop-science, and participated in several national podcasts, radio, and TV shows, where she discussed the issues pertaining to conversational AI. She is fluent in English and French, can professionally translate from Latin, has learned some Sanskrit basics during her bachelor's studies, and is a (poorly) self-taught afficionado of modern Hebrew.

Образац 5.

Изјава о ауторству

Име и презиме аутора *ВАЊА СУБОТИЋ*

Број индекса *ОФ 18-0003*

ИЗЈАВЉУЈЕМ

да је докторска дисертација под насловом

LINGUISTIC COMPETENCE AND THE NEW EMPIRICISM IN PHILOSOPHY AND SCIENCE

ЈЕЗИЧКА КОМПЕТЕНЦИЈА И НОВИ ЕМПИРИЗАМ У ФИЛОЗОФИЈИ И НАУЦИ

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Датум **Потпис аутора**

У Београду, 15.09.2023. _____

Образац 6.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора	*ВАЊА СУБОТИЋ*
Број индекса	*ОФ 18-0003*
Студијски програм	*ФИЛОЗОФИЈА*
Наслов рада	*LINGUISTIC COMPETENCE AND THE NEW EMPIRICISM IN PHILOSOPHY AND SCIENCE*
	ЈЕЗИЧКА КОМПЕТЕНЦИЈА И НОВИ ЕМПИРИЗАМ У ФИЛОЗОФИЈИ И НАУЦИ
Ментор	*МИЉАНА МИЛОЈЕВИЋ*

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Датум **Потпис аутора**

У Београду, 15.09.2023. _____

Образац 7.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

LINGUISTIC COMPETENCE AND THE NEW EMPIRICISM IN PHILOSOPHY AND SCIENCE

ЈЕЗИЧКА КОМПЕТЕНЦИЈА И НОВИ ЕМПИРИЗАМ У ФИЛОЗОФИЈИ И НАУЦИ

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

(6) Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Датум **Потпис аутора**

У Београду, 15.09.2023. _____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.