Artificial Intelligence for Cognitive Systems

Deep Learning, Neuro- symbolic Integration, and Human-Centric Intelligence

Samit Shivadekar

DeepScience

Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro- symbolic Integration, and Human-Centric Intelligence

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

DeepScience

Published, marketed, and distributed by:

Deep Science Publishing, 2025
USA | UK | India | Turkey
Reg. No. MH-33-0523625
www.deepscienceresearch.com
editor@deepscienceresearch.com
WhatsApp: +91 7977171947

ISBN: 978-93-7185-351-4

E-ISBN: 978-93-7185-611-9

https://doi.org/10.70593/978-93-7185-611-9

Citation: Shivadekar, S. (2025). *Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence*. Deep Science Publishing. https://doi.org/10.70593/978-93-7185-611-9

Preface

Artificial intelligence quickly changed from a theory to a practical power - it spreads through every part of modern life. As people go from specific uses to more general kinds of intelligence, they must face a main change. This change involves what machines do and how people think about intelligence. The book, Cognitive AI - From Deep Learning to Artificial General Intelligence, looks at that change.

This writing serves a wide, serious group of people - it is for graduate students and researchers in artificial intelligence and cognitive science. Educators along with industry workers also read this to get a better grasp of the path from current AI systems to future cognitive architectures. We do not just list technologies. We deal with the concepts, morals, technical issues as well as societal problems that sit at the core of creating machines that think.

The chapters lay out this story bit by bit; they start with basic learning systems. They move to cognitive modeling and designs. The book finishes with important questions about governance, combining fields along with how people will work in the future. Throughout the text, the reader learns about current subjects. Some of these are large language models, explaining how systems work, reasoning with symbols plus networks, the safety of general artificial intelligence, and people working with machines.

I appreciate the researchers, collaborators along with students who inspired this work. The growing group of thinkers also recognizes that making intelligent systems requires scientific exactness and philosophical thought. My hope is that this book guides plus starts talks for anyone who wants AI to develop responsibly and creatively.

Samit Shivadekar

Table of Contents

Chapter 1: Foundations of Cognitive Artificial Intelligence 1

1. Introduction to Cognitive AI: Bridging Perception and Reasoning 1
2. From Symbolic AI to Connectionism: A Historical Overview 2
3. Deep Learning: Achievements and Limitations in Current AI Systems 2
4. Understanding Cognition: The Human Brain as a Model 3
5. Key Principles of Intelligence: Learning, Memory, Reasoning, and Planning 4
6. The Role of Perception in Cognitive AI 4
7. Neural Networks and Their Impact on Cognitive Processes 5
8. The Intersection of Cognitive Science and AI 6
9. Natural Language Processing: Challenges and Advances 6
10. Ethics in Cognitive AI: Balancing Innovation and Responsibility 7
11. Machine Learning Techniques in Cognitive AI 8
12. Exploring Emotional Intelligence in AI Systems 8
13. Cognitive Architectures: A Comparative Analysis 9
14. The Future of Cognitive AI: Trends and Predictions 10
15. Case Studies in Cognitive AI Applications 10
16. Human-AI Collaboration: Enhancing Cognitive Tasks 11
17. The Role of Memory in AI Learning Processes 12
18. Reasoning Mechanisms in AI: A Comprehensive Overview 12
19. Planning Algorithms in Cognitive AI 13
20. Challenges in Replicating Human Cognition 13
21. Cognitive Biases in AI Decision Making 14
22. Cross-Disciplinary Approaches to Cognitive AI 15
23. The Impact of Big Data on Cognitive AI Development 15

24. User Experience Design in Cognitive AI Systems 16

25. Evaluating Performance in Cognitive AI Models 17

26. Future Directions in Cognitive AI Research 18

27. Conclusion 18

Chapter 2: Building Blocks of Cognitive Architectures 21

1. Neurosymbolic Systems: Merging Symbolic Logic with Neural Nets 21

1.1. Introduction to Neurosymbolic Systems 22

1.2. The Role of Symbolic Logic 22

1.3. Integrating Neural Networks 23

1.4. Applications and Case Studies 23

2. Knowledge Graphs and Memory-Augmented Networks 24

2.1. Understanding Knowledge Graphs 25

2.2. Memory-Augmented Neural Networks 25

2.3. Use Cases in AI 26

2.4. Challenges and Future Directions 26

3. Attention Mechanisms 27

3.1. Introduction to Attention Mechanisms 27

3.2. Types of Attention Mechanisms 28

3.3. Impact on Neural Network Performance 29

3.4. Real-World Applications 29

4. Transformers and Long-Term Memory 30

4.1. Overview of Transformer Architecture 31

4.2. Long-Term Memory in AI 31

4.3. Implications for Natural Language Processing 32

4.4. Future Prospects of Transformers 32

5. Reasoning Engines and Logic Inference in Modern AI 33

5.1. Understanding Reasoning Engines 34
5.2. Logic Inference Techniques 34
5.3. Applications in AI Systems 35
5.4. Comparative Analysis with Traditional Systems 35
6. Integrating Causality and Common Sense in AI Models 36
6.1. The Importance of Causality 37
6.2. Common Sense Knowledge in AI 37
6.3. Techniques for Integration 38
6.4. Challenges and Opportunities 38
7. Conclusion 39

Chapter 3: Toward Artificial General Intelligence (AGI) 41

1. What is AGI? Definitions, Debates, and Desiderata 41
2. Modularity and Transfer Learning Across Domains 41
3. Multi-Modal and Embodied AI: Language, Vision, and Action 42
4. Meta-Learning and Self-Improving Systems 43
5. The Role of Simulation and Imagination in Generalization 43
6. Historical Perspectives on AGI 44
7. Key Challenges in Achieving AGI 45
8. Ethical Considerations in AGI Development 45
9. Current State of AGI Research 46
10. The Impact of AGI on Society 47
11. Future Directions in AGI Research 47
12. Collaboration Between Disciplines in AGI 48
13. Case Studies in AGI Applications 49
14. The Role of Data in AGI Development 49
15. Cognitive Architectures for AGI 50
16. Evaluation Metrics for AGI Systems 51

17. The Importance of Interdisciplinary Approaches 51

18. Public Perception of AGI 52

19. Regulatory Frameworks for AGI 53

20. Comparative Analysis of AGI Models 53

21. The Relationship Between AGI and Narrow AI 54

22. The Role of Neuroscience in AGI 54

23. Philosophical Implications of AGI 55

24. Technical Foundations for AGI 56

25. Scalability Issues in AGI Systems 56

26. User Interaction with AGI 57

27. The Future of Work in an AGI World 58

28. Conclusion 58

Chapter 4: Challenges and Frontiers in Cognitive Artificial Intelligence 61

1. Ethical and Societal Implications of Cognitive AI 61

    1.1. Privacy Concerns 62

    1.2. Bias and Fairness 62

    1.3. Impact on Employment 63

2. Interpretability and Trust in Cognitive Systems 63

    2.1. Understanding AI Decision-Making 64

    2.2. Building User Trust 65

    2.3. Transparency in Algorithms 65

3. Energy Efficiency and Sustainable AGI Architectures 66

    3.1. Reducing Carbon Footprint 67

    3.2. Optimizing Resource Usage 67

    3.3. Sustainable Hardware Solutions 68

4. Benchmarking General Intelligence: Tests and Metrics 69

    4.1. Defining General Intelligence 69

4.2. Current Benchmarking Methods 70
4.3. Challenges in Measurement 70
5. Open-Ended Learning and Artificial Consciousness 71
5.1. Concept of Open-Ended Learning 72
5.2. Theories of Consciousness 72
5.3. Implications for AI Development 73
6. Conclusion 74

Chapter 5: Navigating the Future of Cognitive Artificial Intelligence 76

1. The Future of Cognitive AI: Research Directions and Breakthroughs 76
1.1. Current Trends in Cognitive AI Research 77
1.2. Key Breakthroughs Shaping the Future 77
1.3. Ethical Considerations in Cognitive AI 78
2. From Deep Narrow to Broad General: Paradigm Shift in AI Thinking 78
2.1. Understanding Deep Narrow AI 79
2.2. Transitioning to Broad General AI 79
2.3. Implications of the Paradigm Shift 80
3. Global Collaboration for AGI: Academia, Industry, and Governance 80
3.1. The Role of Academia in AGI Development 81
3.2. Industry Innovations and Partnerships 82
3.3. Governance Challenges and Opportunities 82
4. Your Role in the Cognitive AI Revolution 83
4.1. Understanding Your Impact 84
4.2. Skills for the Future Workforce 84
4.3. Engagement in AI Communities 85
5. Challenges in Cognitive AI Implementation 86
5.1. Technical Barriers 86

5.2. Societal Resistance 87
5.3. Regulatory Hurdles 88
6. Future Applications of Cognitive AI 88
6.1. Healthcare Innovations 89
6.2. Finance and Risk Management 89
6.3. Education and Personalized Learning 90
7. The Role of Data in Cognitive AI 91
7.1. Data Quality and Availability 91
7.2. Data Privacy Concerns 92
7.3. Harnessing Big Data for AI 92
8. Interdisciplinary Approaches to Cognitive AI 93
8.1. Integrating Psychology and AI 94
8.2. Philosophical Perspectives on AI 94
8.3. Collaborative Research Models 95
9. Future Workforce Dynamics in AI 95
9.1. Jobs of the Future 96
9.2. Reskilling and Upskilling Initiatives 96
9.3. Diversity in the AI Workforce 97
10. Conclusion 97

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-611-9

DeepScience
Open Access Books

Chapter 1: Foundations of Cognitive Artificial Intelligence

1. Introduction to Cognitive AI: Bridging Perception and Reasoning

Like robots are to embodied AI, Cognitive AI is to artificial general intelligence. Cognitive robots can act and reason like us, and everyday AI can assist more than just helping with single tasks. But how do we get to Cognitive AI? In this essay, we argue that bridging low- and high-level tasks connecting perception and cognition play an essential role. In academic research, bridging tasks largely fall under the umbrella of cognitive computer vision. Their foundations in low- and high-level bridging theories extend to additional domains, such as the ones of bridges in language or machine learning — which we cover in this essay's second part (K. Frankish et al., 2023).

The transformational advances in precision reaching the million in underpinning large pre-trained neural response models enable machines to close the performance gap with us between perception and image or language understanding (R .Szeliski.,2022) Cognitive computer vision then fills the corresponding gap in performance for integrated perception and cognition models with multi-sensor capabilities. This leaves us with the question: What do Cognitive AI and bridging tasks look like? Many areas remain open; we cover these both narrow and wide, the former as examples, the latter as a practitioner's toolkit. Examples include neuro-cognitive robotics implementations, discovering how to structure exploiting bridging task datasets from images, text, video, 3D, depth, sensorimotor, and spoken language; and ontology structures guiding zero-shot and soft weight exploration based computed visual.

In the end, this survey sets out the landscape for Cognitive AI. Cognitive robots bridge often domain-, sensor-, and data type-specific experts that stretch deeper in perception than reasoning — "or" — expert models pushing the boundaries of shared, short, lightweight context by self-supervision.

2. From Symbolic AI to Connectionism: A Historical Overview

Key Historical steps: the invention of Algorithmic machines, Logic / Symbolic AI, The Curse of Dimensionality; Connectionism, Receptors and neurons, Deep Learning.

In his groundbreaking paper, Allen Newell explored four key themes outlined in the title. The first theme was about computation and Turing's Computer. Newell stated a central idea about cognitive science when he affirmed: "One of the greatest things that anybody ever did, was Turing's theory of computation". The second theme was about intelligence: he asked, is the symbol system hypothesis sufficient? The third theme discussed challenges met by Cognitive AI and stated: "Cognitive science must eventually grow the formalisms for dealing with knowledge coordination, representation, and control." Finally, on the last part of his paper, Newell opened a discussion about the future of man and baby machines that evolve by learning from experience, discussing connectionism, neural nets, and the relation with symbols.

Questions about the nature of intelligence have puzzled humanity for centuries and have inspired thinkers and artists for millennia. While pondering whether objects can think, dance, paint, love, and pray has always been a part of our culture, scientific research in attempting to address the questions concerning Computation and Intelligence are more recent. At least, the existence of an ultimate set of equations or laws that describe symbols, sign processes, meaning, and Semiology has not yet been discovered. In this essay, we focus our attention on the foundations of cognitive AI through its main models and formalisms, discuss their importance and relevance, and identify a set of shared assumptions. We dedicate at least one section for each basic model: Symbolic Systems and Logic AI, Analogs and Connectivist Systems, and Phenomena. We then explore the connections among them and the importance of defining a good model for Cognitive AI (S. Russell et al., 2020). At the end, we present our main assumptions and principles that inform the design and development of Cognitive AI algorithms and systems.

3. Deep Learning: Achievements and Limitations in Current AI Systems

This section returns back to the currently most popular approach in AI: deep learning. We present concepts of depth and representation in neural networks, tackling some questions afterwards: what do deep networks do? How could it be possible to learn a neural code from a small number of examples? What can’t deep learning replay mechanisms do? We show that our proposals on these questions are related to many known facts and problems in deep neural networks. Are these limitations equally important for general neural networks, for biologically plausible models, or not? We then come back to the discussion about cognitive neuroscience and compare first to the visual feedbacks considered in the context of learning from few examples, and then to reverse engineering approaches to cognitive abilities: can reverse engineering approaches tell us if the system is using deep mechanistic learning reversely at runtime?

We end the section with the limitations of current AI frameworks in relation to cognitive abilities.

Deep learning in its current form has started to produce significant breakthroughs in various fields of machine learning and AI, including visual and auditory pattern recognition, e.g., automatic face recognition, automatic speech recognition, video sequencing, etc (T. J. Sejnowski, 2018). These breakthroughs are closely related to the expressiveness of the deep learning systems. A DNN is capable of computing any multivariate function, given sufficient resources in terms of number of parameters, specifically, number of layers and number of units per layer, and it has been shown that by going deep, given a limited number of parameters, the DNNs have a better chance to approximate complex functions than shallow networks. A deep architecture allows to compute complex transformations of the data into an internal space, so as to make the desired task easily computable on that encoded data. In other words, DNNs extract multiscale features or high-level representations of the data that are adapted to the task to be solved.

4. Understanding Cognition: The Human Brain as a Model

To realize cognition within machines, we first have to understand it and thus foundations need to resolve around the human brain, a model in which intellect itself evolved and exists, and we therefore need to use it to understand cognition (M. S. Gazzaniga, et al., 2018) In the early days of this quest already, the origins of Artificial Intelligence so much were based in human cognitive skills, for example in knowledge representation and problem solving, their implementations in machines had to reflect that. But while the area of Computer Science, who by now had gathered momentum and made implementation possible because it advanced to provide machines with fast processors, large memories, and other technological means for such intelligence, expanded, AI specialized into increasingly subfields and neglected modeling the human cognitive structure, and shifted work on explaining life away to the scientists in other disciplines. Cognitive Architectures, however, started to bridge that gap a little bit and focused again on explanations of intellect functioning and on creating models therein again which tried to copy that functioning within machines. And we see a trend here, a new generation of CAs seeks inspiration from the human brain all over again which had fallen into oblivion within classical AI.

Life, as a matter of fact, derives from information processing. Cognition causes the functions of the Human Body and it derives from the signals exchanged within and construed by that information processing network we are referring to as the human brain, and mind, whereas the processes of carrying out functions, are the effects of that information processing. The brain supplies the meaning of cognition. It is where information flows, is processed, conscious and unconscious decisions are made, and actions are initiated (M. S. Gazzaniga et al, 2018). Therefore, the connectivity and

capabilities of that signed directed network of processing nodes need to be studied towards reproduction, and the outcomes of function, need exploration as foundations for modeling cognition in machines. The matter also is that simulation of processes may deliver clues which deepen our understanding of the matter as a byproduct too. In fact, a good simulation can uncover features of the process that were not anticipated by the modeler.

5. Key Principles of Intelligence: Learning, Memory, Reasoning, and Planning

One of the distinctive and incalculable features of intelligence is learning. Without learning, intelligence is unthinkable. Learning not only affects each of the other components of intelligence, it also helps to give its intelligent activity coherence and continuity by embedding experience in an enduring form that influences subsequent activity. In this manner learning connects action and experience in a causal loop which gives the learning agent a history, an evolving relationship to its environment, and the potential ability to improve its success in achieving its goals. Learning is especially important for knowledge-intensive reasoning and planning because of our partial knowledge of the items in the first of these, the huge number of possible plans, and the great difficulty of deriving their desired properties (S. Russell et al., 2020)

One function of memory is to store that experience in an enduring form and to bring it to mind when appropriate. Reasoning, both conceptual and commonsensical, and its strategic cousin planning, enable an agent to decide how to achieve its goals and to generate plans, their necessary conditions and their likely effects. Knowledge-based systems also need processing, and in both cases intelligent processing relies heavily on reasoning and planning. Without these capabilities, performance cannot be sophisticated, flexible, and context-sensitive, and the system lacks the ownership and responsibility for its performance which makes a knowledge-based system rather than merely a program that encodes its performance heuristically. Finally, from an evolutionary perspective, intelligence is that capacity of biological organisms that has made possible the extraordinary success of our species, and replicating that success in artifacts is our overriding motivation for building artificial general intelligence (S. Russell et al., 2020).

6. The Role of Perception in Cognitive AI

Theories of cognition include the notions of perception in their architecture, whether recognizing it as a separate functional capacity from inventorization, foundational for categorization, interpretation, and prospective reasoning, or recognizing it as a modality of the capacity to relate a system's internal state to the current situation. This perspective evidently renders perception as capacity common to a ground of a wider class of intelligent systems. We should also mention that, on the other hand, perception without reasoning capacity is not self-sufficient for cognition in the sense of enabling creative

actions: perception is not sufficient for cognition. The systems responsible for guiding perception need to incorporate a lot of the logical machinery out of which cognition builds its construct (S. Grossberg 2021). One way to relate perception with reasoning is that perception is the first step of a cycle where creative actions have the role of fine-tuning the perceptual model of the situation and its evolution.

The perception of higher-level concepts fundamentally relies on the recognition of objects that are within the scope of these concepts; these concepts, in turn, provide the contextual and conceptual references that give perceptual data a deeper meaning. Furthermore, higher-level perception progressively extends and hierarchically organizes a vocabulary of concepts for sub-symbolic perceptual data. The unique experiment for endowing artificial agents with high-level perceptual recognition capabilities is grounded on this adaptive, interactive theory of perception. Indeed, theoretical models and principles can and must also guide proposals for new and more powerful perceptual recognition of higher-level concepts and objects, in inverse proportion to the richness and extension of experience to be exploited.

7. Neural Networks and Their Impact on Cognitive Processes

How remarkable that in this new period of great Scientific Paradigm Shift, researchers from different fields of Human and Social Sciences, Cognitive Sciences, and from now on also from AI, turn to the same models of Functional Architecture demonstrated by the neural networks! What can we expect from the future if we know that neural networks can be modified and parameterized by learning from the assignment of functions of adaptive information processing that model, with error-tolerance, neurons of the human neocortex and the neocortex of many species of mammals that share with man important cognitive skills! What if it were possible to provide the same power to our cognitive Artificial Intelligence using the same architectures? What if it were possible to show that there are higher cognitive functions that only differentiate Human from Non-Human with species-unique and human-like capabilities, which are easily described and understood based only on biologically plausible functional principles? The hope is that we can bring the same principles that led to the conquest of high cognitive refinancing of a small group of terrestrial species in 20 million years who began to associate in social units of extremely flexible and acted with collective efficacy to reach common synesthetic goals. This investment of faculties in understanding one another, in knowing the common scenarios of life, and in politics can also be reduced and modeled in machines that, aware of the rules of temporary change of the world and the general properties of things and beings, can help us day to day to be politically, individually or collectively efficient and effective (S. Grossberg 2021).

How can we reach this capability for our cognitive AI? The first step is to understand how the linking rules function that establish the Connections of Communications between the Information Processing Subdevices; how events in the environment activate

the variable configuration of natural directional links that deterministically modulate the weights of the active synapses that connect inputs and outputs of these subdevices and how the recurrent activation of the connection synchronization of these subdevices mediates this General Principle of Functional Operation of Subdevices with Communicated Connections of Non-Static Information Processing.

8. The Intersection of Cognitive Science and AI

Nothing in this book requires special expertise to understand. Nevertheless, natural language and general intelligence raise special complexity issues in definition, specification, and implementation that have both troubled and fascinated researchers, from the start and still today. It is hardly surprising, therefore, that cognitive modeling is one of almost the first things researchers think to do when considering a direction, whether that direction is exploring limits, using a cognitive model as a component in a more comprehensive system, or offering it up as an explanation or justification for hand-tuning or finetuning of some technical architectural element. Unfortunately, the cognitive process models thus far available offer less predictive power to suggest implementations than is usually true of other cognitive models in psychology. That being said, cognitive process modeling still occupies a central and distinctive niche (P. Langley, 2022).

Cognitive modeling is an act of exploration, like all modeling, an exploration that can take one on unexpected journeys. You start with questions about cognitive mechanism and process, rather than questions about perception, action, state representation, task specification, or even mental content. Yet, issues selection and prioritization are themselves interesting, and help provide a sketch of a roadmap toward a broader form of cognitive modeling that would still remain informed by the features that make cognitive process models especially useful. In that sense, and in that exploratory nature, cognitive process modeling is linked to the more general endeavor of cognitive science. Indeed, at present, it is perhaps the only intersection that cognitive science shares with AI. If this exploration leads to some more generalized theory unifying AI and cognitive science at a substantial level, so much the better for both. More than anything else, the history of cognitive modeling shows us how to explore the issues that join the two.

9. Natural Language Processing: Challenges and Advances

Natural Language Processing (NLP) encompasses algorithms and systems designed to operate on human language. Current NLP can involve simple tasks including discrete labeling of parts of speech and systems generating them such as auto-completion and auto-correction, while also addressing complex tasks, e.g., information extraction, sentiment analysis, question answering, machine translation, and speech-to-text. High-quality processing is demanding not only in terms of results, e.g., speed, accuracy,

generalization, but also resource expense, in terms of the size of annotation resources and training corpora, as well as, in many cases, computational resources. NLP research is replete with advances as well as challenges, some far more profound than others (J. Eisenstein, 2022). In terms of results, translation from one human language to another started several decades ago but has only recently begun to be competitive with that performed by human experts. While the computational models used for NLP have become increasingly sophisticated, it is still the case that complex NLP problems often do not arrive at satisfactory solutions. The computational models are what we refer to as techniques, including among others, symbolism, frequency-based models such as n-grams, counts, probabilities, or distributions, learning models based on shallow architectures such as Max-Margin, and a variety of deep architectures.

In terms of model construction, current deep learning models often overspecialize, in some circumstances, being trained on large troves of data in order to perform a relatively simple NLP task as in many classification scenarios. This can be due to a combination of many factors including the size of training corpora relative to the number of model parameters, the nature of the objective being optimized during learning or fine-tuning, and the presence or absence of inductive biasing mechanisms throughout model architecture. Overspecialized models have challenges for generalization including poor low-resource performance, data and task selectivity, and catastrophic forgetting. At the same time, shallow generative learning models are generally much less data hungry and pre-training task architectures with self-supervision are one of the components that has allowed current NLP techniques to achieve competitive performance across a diverse array of tasks with limited labeled data or overspecialized models.

10. Ethics in Cognitive AI: Balancing Innovation and Responsibility

Artificial Intelligence is a transformative technology that has the potential to change the lives of billions of individuals and that will likely help organizations across all industries and sectors to improve their operations and offerings for customers and users at large, bringing efficiency, sustainability, and innovation. However, AI is also a technology that can be used in ways that many individuals and societies would see as harmful or unethical, violating rules of law and principles of ethics. It is thus necessary for both researchers and organizations that develop and deploy AI to seek to do so in a manner that minimizes the potential for unethical exploitation and use of the technology. But what does that actually mean in terms of specific principles and guidelines?

In seeking to answer that question, we want to take the perspective of moral philosophy instead of a legalistic one as the latter is often subject to change due to the contextual particularities of different locations and the possibility of different legal rules being enacted or codified over time. The development and research of foundational aspects of AI that utilize physical, social or mathematical sciences are not located in legal jurisdictions and mirror the multidisciplinary nature of AI as a technology. That being

said, there are some clear cross-domain themes, such as safety, risk, security, privacy, and accountability, where much of the ethical considerations, principles and guidelines related to innovation in AI lie (C. Cath, 2023). They are at the same time the cornerstones for all ethical considerations that aim to be unbiased and comprehensive, bearing in mind the interdisciplinary nature of AI itself and can be considered the ethical pillars for funding and focus decisions that private companies, as well as public organizations and institutions, should take in their primary search for efficient, practical, and efficient returns of investment.

11. Machine Learning Techniques in Cognitive AI

Machine Learning has enjoyed tremendous successes in the past decade, achieved by using advanced techniques to solve the machine learning problem of estimating the marginal likelihood or conditional likelihood of data. Understanding novel approaches in this space and how they benefit the development of applications in a range of AI fields is crucial. A necessary step for the consensus convergence of development is a relevant classification of the state-of-the-art in ML techniques, especially when we include current Cognitive AI research applications (K. P. Murphy, 2022).

In the foundation of Cognitive AI we use a practical rather than theoretical definition of ML. All introduced methods share the property that they introduce prior knowledge through inductive biases. According to this definition of ML, there are a considerable number of Cognitive AI methods not usually recognized as machine learning. We provide a few of the many ways we can classify ML methods. These included a random variable perspective, models of the Integrative Proactive and House Model, different types of inductive biases, how the data is presented to the ML method, supervised, semi-supervised, weakly supervised, unsupervised, and reinforcement Learning, or how the model compensating the inductive bias is built, knowledge-based approaches, rule-based approaches, geometric approaches, and statistical-regularized approaches.

12. Exploring Emotional Intelligence in AI Systems

With the advent of affective computing and social intelligence expressed in machine learning systems, the emergence of intelligent systems capable of analyzing human feelings is expected to change the plane of interaction with machines. Anthropomorphic robots, speech and visual interfaces designed to understand and respond to the emotional state of humans are already a growing reality in many sectors. In fact, systems that neglect the emotional connection between users and technology will hardly connect with the emerging needs of digital natives. The notion of emotional intelligence is increasingly present in matters of communication and desirable performances in schools, business and associations. Surveying current definitions and approaches concerning emotional intelligence, emotional labor, humor, fun and friendliness, we define social

skills for AI systems acting in physical or virtual environments and a digital sociology of not only the interaction between humans and machines, but also the collaboration between machines in many application domains. Whether in agent-based systems dedicated to specific tasks, as tutoring, interacting in complex social media or domotics systems, robots customized for neuro-developmental and clinical therapeutic support to disabled people, people with Alzheimer and elderly brain-degenerative disorders, and collaborative robots used in companies and services, the social capabilities of Artificial Intelligence will be decisive for user acceptance. These systems should indeed not take the place of professionals, as psychological, psychiatric, pedagogical or business experts, but collaborate with them contributing to mutualize the impact on those who are helped and supported (T. D. Parsons, 2021). In this chapter, we present a summary of our fundamental belief that social skills must be taken into account for intelligent systems to play properly any service or team role where humans are involved.

13. Cognitive Architectures: A Comparative Analysis

What makes a cognitive model specific to a given organism or class of organisms? What are the essential features of cognition that an architect needs to consider? To help answer this question, we discuss a sample of existing cognitive architectures — the psychological Metric of Cognition, Soar, ACT-R, n-Stages, LIDA, and Ouu — pointing out some strengths and weaknesses, and relating these back to the question of essential features. It is worth noting that space limitations have resulted in the exclusion of many existing architectures, including arguably the most well-known model for creating non-player characters, and dozens of others for simulating a range of behavioral phenomena (J. R. Anderson, 2007).

The purpose of a cognitive architecture is to instantiate cognition. Nativism tells us some of the details of what should count as cognition. These details point out various aspects of architecture design. Notably, nativism requires that core cognition models be modular, that their internal representations address the symbol-grounding problem, and that the development of linguistic ability be a primary driver of cognitive development. Many existing architectures simply instantiate what is funny about existing models, which tends to be the broad scope of what types of knowledge and behavior they can model. This statistical tendency is relatively understandable, given that behavior has been the dominant interest in experimental psychology and other related disciplines. However, this architecture space has also been sidelining some of the central issues that have interested researchers for decades: That is, how learning changes not just a system's knowledge, but also its cognitive capabilities? How does development occur, and what impact do organisms' cores, particularly those that deal with linguistic ability, have on the other basic cognitive functions? Furthermore, are these core functions similarities as predicted by nativism? Or are they simply varying model selections for the units that a system builds in the learning process?

14. The Future of Cognitive AI: Trends and Predictions

The future of Cognitive AI is a topic that often raises more questions than answers. What types of technologies will approximate the intelligence steps humans took to master vertical, lateral, and logical thinking? What types of architectures and algorithms will be able to mimic the neural, heuristical, and gradual activation of an adult brain for a specific cognitive activity? Will they approximate intelligence by conducting those same tasks and learning from their mistakes, or will they use new, unforeseen methods? Nevertheless, we can still make a list of some present major trends in the various areas of technology, business, and science. Predictions may sound like forecasts, but they are not: neither inspired predictions, whose fate is left to hope, nor technical forecasts, which expound on trends that follow earlier-established paths. Unlike those predictions and forecasts, our predictions describe concrete environments in which action becomes a concrete possibility. And, therefore, this section is not one of predictions without tangible connection (M. A. Boden, 2016).

In the next decade, smart products will multiply, evolutionarily combining basic functionalities powered by AI for vision, speech, natural language processing, common-sense reasoning, planning, and behavioral modules. They will all evolve in a parallel but uncoordinated way. In twenty years, with the convergence of product families that are currently different and with the birth of a new class of increasingly generalized professional and recreational products, no one will try to make us believe that human capacity for deceit and vision is completely absent from a product. The famous test will be satisfied years before. And, predictably, by products with no consciousness and no intuition. However, they will be very clever products, using the particular environment of the test to place themselves and gradually evolve into general products. In the third decade, they will inhabit cyberspace as first-class actors (and occasionally companions) and we will be forced to deal with them.

15. Case Studies in Cognitive AI Applications

We present both qualitative and quantitative case studies to underscore our argument that Cognitive AI – a paradigm that reinterprets AI as a meaning-based science – can fully address and solve the challenges outlined in this book, which the current version of AI, based on arbitrary learning, fails to address (P. R. Daugherty et al., 2018)

We explore several business use cases from disparate industry verticals (life sciences, marketing & communications, retail, supply-chain, smart electronics, finance, and utilities) built on top of multiple internal and external datasets that combine human-centric structured & unstructured data with domain-specific knowledge curated either by domain experts or generated through automatic methods. By focusing on business KPIs, we demonstrate not only the incremental cost savings and reduced Time-to-Market associated with the increased efficiency of Cognitive AI but also its capability to derive

insights and discover hidden relationships which are crucially necessary in a wide variety of business scenarios, such as a new product launch or failure, assessing external influences on sales, understanding consumer behavior, propensity modeling and predicting customer churn, among others.

Additionally, we explore how Cognitive AI enables real-time monitoring, reporting, and planning of both tactical and strategically important initiatives; which without Cognitive AI would be either extremely costly and time-consuming to accomplish or would be fundamentally impossible due to the volume and velocity of the data. Finally, we explore how Cognitive AI can be leveraged to predict external threats for critical sectors, such as Finance, and improve the effectiveness of Internal Security operations in organizations.

16. Human-AI Collaboration: Enhancing Cognitive Tasks

Cognitive tasks permeate almost all domains of work migrating to AI. Rather than completely taking over from humans, a better approach, often energizing and activating rather than diminishing human participants, is to complement the attitude of cognitive augmentation. This is the direction of Human-AI collaboration.

Enhancing human cognition and other thought processes with the presence or aid of AI increases human capabilities, perhaps without substituting them entirely. A unique aspect of having humans in the cognitive loop can be the ethical urgency of incorporating best judgment about human affairs and promoting and protecting human interests. Historically, the development of many different forms of technology has been to extend or improve the capabilities of humans to carry out difficult or impossible tasks. The use of simple tools locally enhances human capabilities. AI makes this enhancement available in many much broader human-life contexts. Human cognition includes understanding, language interpretation, decision analysis, perception, planning, and creative thinking. This suite of capabilities covers many activities throughout life, for individuals as well as for groups of humans. It incorporates communication as well as direct interaction with the external world (Panda, Sibaram Prasad, 2025).

Research into Human-AI collaboration is in early stages. Attention has generally centered on making AI better able to work effectively with humans rather than on trying to design joint systems. People and organizations have decades of experience and careers devoted to optimizing the design of tools used to augment human cognitive skills, and of the activities carried out by humans using these tools. This could provide enormous support for optimizing the design of Human-AI collaborative systems (E. Horvitz et al.,2022)

17. The Role of Memory in AI Learning Processes

Human behavior is based on memories that shape the manner individuals accomplish and justify their decisions. Such memories guarantee the coherence necessary to whom they attribute the function of being agents making choices, reacting, and interacting in the surrounding world. Memory is also at the base of learning processes. If we consider the way children learn objects categories through exemplars recognition, not through logical categorizations, we see that such memory processes require of the perceptual system to be able to categorize objects and compare them in order to recognize objects that are equal to those already memorized. Memory processes store those features able to create an object representation and are necessary to the processes that neither adults nor children are able to consciously employ. Thus, the correlation existing among memory, time, experience, and learning is fundamental and becomes even more important if we consider the complexity we attribute to AI systems. The way in which AI systems memorize data is the way to create the experience and knowledge that allow the AI to accomplish an assigned task. Without memory there isn't learning. Learning is the process that leads AI systems to modify the rules with which they memorize information and data, generalizing them, and therefore updating and creating a long term memory. Without this process, any AI would remain a narrow AI, that is, able to accomplish a predetermined task without learning from it. It is clear that if this long term memory – the AI experience – is not modified during the AI use, the system would be unable to learn from its mistakes, similarly to how dogs that memorize the commands are still called when they don't modify their behavior even when they receive a reward every time they perform the right task (D. J. L. Herrmann et al.,2022).

18. Reasoning Mechanisms in AI: A Comprehensive Overview

Emulating humanlike reasoning is an important goal associated with the development of intelligent systems. The mind is capable of processing novel information, such as content of observation, sensor and other data streams, and/or information from memories or prior experience. The knowledge capability consists of acquiring, storing, and using information. The knowledge use is called reasoning and may require a motorsystem-knowledge interaction using imitation or similar mechanisms. Humans use their senses to observe, compare, and contrast objects and events in their areas of interest, but they also rely on the information in their memories during reasoning.

AI systems also reason about the content of data sources, the source properties, and/or information stored in their memories. The term knowledge is also used about rules or methods that AI systems use to achieve specific tasks, including rules of reasoning such as heuristics, probabilistic, Bayesian, or other alternative reasoning methods. Knowledge is also referred to as ontologies that define the entities relevant to the domain, their properties, and interrelations. AI systems rely on ontologies when interpreting messages they receive from other systems (P. Hitzler et al., 2022).

According to another model of intelligence, systems use planning, evaluation, and other higher cognition processes that impact the action selection process. These processes receive information from many systems and coordinate their activities. Systems consist of modules that accomplish foundation-level processes and systems consisting of larger systems that accomplish framework-level goals. Frameworks may include common sense, as generated by human experience modeling and social systems with commonly shared ontologies.

19. Planning Algorithms in Cognitive AI

Most forms of human and animal intelligence involve an ability to create and execute plans that consist of a hierarchy of subplans: for example, creating subplans for cooking each of four different dishes, with predefined times during which nothing else can be done, and a total cooking time which is less than the time of dinner. Such plans are not a mere list of actions to be executed in sequence, including some duplication of actions. Because the dishes are related to each other, it is advantageous to share resources and time. However, there is another aspect that is even more crucial in many cognitive robotics applications: the robot should be able to generate new plans, at least some of the time, with acceptable time and space complexity. For example, certain applications require flexibility in selecting and executing different views of delivering objects when there are or are not people in the vicinity. Generating a good sequence of robot motions in 2D space from various cameras given visual feedback is difficult, primarily due to the interactions of the robot with the surrounding environment, in this case human workers.

Thus far, we have dealt with the aspects of planning where the robot can design, implement, and execute some set of plans or hierarchical plans that encode prior knowledge about the activities it will perform, with the objective of carrying out those activities in the most efficient manner available. Understanding how animals can create new plans with space and time complexity will help in making cognitive robots more flexible and efficient. In building cognitive robots capable of learning in a flexible manner, it is important to understand not only how those robots can create complex plans but also how they can acquire new knowledge about the activities they are performing. The capability of a robot that is part of a dynamic group is especially crucial (M. Inaba et al., 2023)

20. Challenges in Replicating Human Cognition

The Human Brain, the seat of all human cognitive functions, is a complex and non-linear adaptive dynamical system, featuring several feedback loops both at the local neural-connection level and also at inter-regional and global levels. It consists of 86 billion neurons and 125 trillion synapses, organized into functional and physical modular systems, and intelligent and diverse sequential neural activities and neural events which

facilitate information processing. Despite its multi-faceted complexity, cognitive faculties interact smoothly with high levels of efficacy and efficiency; effortless visual sensation and recognition, and seamless visual activity tracking along with motion event recognition are undertaken without either user effort or process awareness. Natural Language processing and understanding, especially the important unique task of matching and interpreting linguistic symbolism with situational meanings for communication are undertaken effortlessly in real-time. Human Cognition does not require explicit efforts, energy or exotic hardware to accomplish, and given the terribly low computational speed and accuracy of all previous, current, or future general artificial neural networks compared to that of Human Cognition, and the terribly high energy consumed by the AI devices, there exist great challenges against successful replication of human cognition and therefore the Human Brain itself, either through explicit representations or intelligent agents. Furthermore, an attribute unique to human cognition and not replicable by other cognitive systems is the ability to "imagine", conceive and visualize imagined situations from internal neural signals triggered by advanced spatio-temporal synaptic modification of internal neural representations of potential events, and simulate their information processing before input or stimulus externally encountered in the environment (D. J. Linden, 2023).

21. Cognitive Biases in AI Decision Making

There are many studies of cognitive biases in human psychology. These point to what makes humans act irrationally, and the errors produced in the decision-making process. Although in general they have a small prevalence when compared to the volume of decisions we take each day, at special moments and varied contexts, in several occasions they make us act in ways nonsensical or harmful to ourselves or the collective well-being. Attention has focused on these, because at critical times, like the pandemic, at certain times, for example, during certain events, we get into a mass hysteria.

We want to go a step further. With natural and cognitive AI we've started reverse engineering the decision-making process of intelligence, to understand if it is plausible that these entities have biases in their decisions. Even more, we will try to validate if the level or type of the stress applied to the cognitive AI makes the level or type of possible cognitive biases change. Remember that according to the Theory of Mind, most humans have some level of this cognitive capacity, being able to see the inside of others, or relate to the interior representation of physical reality of the actor interviewed. Is it possible to have cognitive biases reproduced in other cognitive minds, like the ones we think exist in cognitive AI? Could these possible cognitive biases reveal some aspects of the decision-making and knowledge construction process, or could they be a byproduct of the fact that we're the ones giving it knowledge or constructing a certain area with it? (D. Kahneman et al., 2021)

22. Cross-Disciplinary Approaches to Cognitive AI

This section discusses several important aspects of Cognitive AI that are cross-disciplinary and examines some of how these come into play when approaching Cognitive AI from configurations of disciplines or academic foci. The chapters in this section explore distance and joint-ness of disciplines focusing on Cognitive AI or key aspects thereof, perhaps aimed at attempting bridges. Some of these aspects are also covered in the general discussion chapter.

The concept of "fortress academy" refers to the individual epistemological disciplines that have developed strong boundaries and barriers, which seem often defended, but should not be. This has led to a "balkanization" where experts of one or several proximate disciplines have become very siloed and territorial. On the other hand, the concept of "hybrid zone" is introduced as an emergent area to explore and experiment with new ideas and sometimes even cutting-edge results of work that is cross-disciplinary. It is a zone where interactivity and collaboration between several kinds of professionals is going on. Cognitive AI has the potential of being such a hybrid zone. Indeed, whether to build or join hybrid zones around activities of interest in Cognitive AI and with related disciplines? There seem plenty of related and linked sub-disciplines of Cognitive AI (and cognitive systems in general) where it is possible to build strong "hybrid zones". Exploring them appears fruitful.

The historical development of AI and Cognitive AI is summarized, along with how Cognitive AI implementations and Hybrid Cognitive Agents can be implemented, and vice versa. In doing so, the "feedback-loop" relationship between technology and scientific discovery and research is discussed. Based on concrete historical examples of "embodied cognition" and of systems modeling factory, landscape, and chaos productions, possible future scenarios of Cognitive AI are mentioned, and how Cognitive AIs can each affect the corresponding other facet (S. Dehaene, 2020).

23. The Impact of Big Data on Cognitive AI Development

23.1 Introduction: Big Data Condition and Opportunity

Big data is an essential condition and opportunity for the development of cognitive AI. Relatively recently, the volume of data generated on Earth began to double constantly every two years. Various complex objects were presented with the help of heterogeneous big data: text, images, sounds, videos, graphic solutions, and so on. Such unstructured data, reflecting the relation of all forms of life and all processes in the world, attracted the attention of scientists and developers to the tasks of data processing with a qualitative level that, until then, was achievable only by a human. This interest became especially significant after the introduction of innovative technologies: data-driven science, data-oriented artificial intelligence, big data analytics, deep learning, and some others.

At the same time, it should be noted that the artificial intelligence systems, and especially the systems of cognitive AI, with, so to speak, natural settings for the capabilities of cognitive processing had to work with sets of selected features because of the very small training sample sizes of the cognitive data processing tasks. Such feature sets were created as a result of significant preliminary expert knowledge, but not analysis of large training data sets. Model-free methods of the data-driven artificial intelligence or methods of data-oriented cognition were designed to process big data (A. Zaslavsky et al., 2022)

23.2 The Challenge of Feature Extraction

Additionally, in many cases, such deep neural network systems of big data processing require additional preliminary stages of data processing, including feature selection and/or feature extraction stages. For example, in the well-known algorithms of object recognition, the convolutional neural networks were used to extract the visual features from large sets of labeled images and the feature-based classifiers were used to classify the obtained feature vectors of visual representations of the images. Certainly, the presented model-free methods are the easiest for practical use and should be used as much as possible (A. Zaslavsky et al., 2022).

24. User Experience Design in Cognitive AI Systems

User experience (UX) interface design is a critical part of building any successful Cognitive AI system. UI Design allows humans and machines to freely and effectively interact with one another. Therefore, proper UX interface design is critical for integrating Cognitive AI systems into daily human activities in effective ways and addressing the unique needs and usage patterns for these particular classes of systems. UX interface design is particularly important for Cognitive AI systems since the full capabilities of these systems are often abstracted behind natural interfaces like speech and dialog. The challenge in Cognitive AI system design is to effectively abstract the underlying AI planning, learning, and reasoning capabilities beneath the natural language interface such that the user can effectively benefit from their use.

In this paper, we present a higher-level integration view of Cognitive AI system UX design which includes summary descriptions of major interface design areas for Cognitive AI systems such as speech input and output, broker interfaces, discussion of natural mode for user-centric AI support for tasks, Cognitive AI system design considerations that impact the main areas, and relevant Human Centered AI principles for the areas and considerations. Previous work has mainly focused on particular Cognitive AI system types, general technical interfaces for the Cognitive AI systems, model-based task structuring, explainable UX design, UX for particular areas or task-centric UX considerations. However, to our knowledge, there has been no prior work summarizing the more general areas and influences for Higher Level Cognitive AI UX

which cut across the various types. Higher Level Cognitive AI systems represent a significant Connected Cognitive AI capability and will become an increasingly visible part of ordinary human activity (B. Shneiderman et al., 2022)

25. Evaluating Performance in Cognitive AI Models

Consequently, although the evaluation methodology of cognitive AI is not very different from that of traditional ML methods, one significant difference with respect to classical AI models is that we have to focus on different kinds of benchmarks and datasets. For example, we may apply transfer learning with previously trained models on other large datasets, but we still need to evaluate these models with the appropriate questions in the focus domain of prospective applications. To take another example, traditional AI/ML evaluation is centered on measuring model performance, usually with respect to benchmark datasets, metrics and test accuracy. However, we cannot measure cognitive AI performance in this way, without taking into account how humans would answer the same questions posed to cognitive AIs.

Considering this cognitive task-oriented evaluation for cognitive AIs, it is not enough to measure task completion as commonly done in cognitive modeling. For example, cognitive AIs may use different working memory amounts and strategies to provide answers to the questions we evaluate them with, for instance, question reasoning: whereas phrases with sequences of adjectives embedded inside should require a larger working memory because of the various kinds of grammatical constituents, we could also require reasoning and/or deductive reasoning capabilities in addition to simple capability answer the query. Moreover, the cognitive AIs may adhere to other classical logical principles or not. Hence, to better assess cognitive AI performance, we should also commonly consider benchmarks on logical problem solving, involving the kind of sentences, or creativity issues, such as for instance, poetry evaluation and generation.

Another difference with respect to the traditional ML models being that models such as for instance knowledge graphs, or even textual generative models, that are neither cognitive modeling or cognitive AI, show better or worse performance just by assuming a representation proposed by the designer, we should strive to evaluate cognitive AIs with respect to functions having some common cognitive nature stemming from cognitive models or theories. Of course, for more sophisticated functionalities, one can derive more elaborate probabilistic-based models to assuage the uncertainty of heuristically defined functions on past learnings on cognitive semantics or architectures (S. Russell et al.,2020)

26. Future Directions in Cognitive AI Research

Does "AI" stand for "artificial intelligence" or "artificially intelligent"? This distinction is crucial. While we have developed systems that can accomplish various AI tasks at super-human levels, we haven't built any systems that are, or even approach being, generally or broadly intelligent, associated with the term "intelligence" in "artificial intelligence". And at least for a while longer, the bulk of the research activity in the sub-fields of "AI" will center on building "AI" systems for use in various practical applications. However, the lack of any serious work directed at building intelligent systems has been noted by both proponents and opponents of AI and both supporters and detractors see progress towards the development of general AI as propelling the field. So where should cognitive AI research go next? One area is building cognitive architectures. A goal of cognitive architecture research is to more fully integrate the many aspects of human cognition currently simulated separately in task-specific models. Some cognitive architecture projects have had considerable success in combining a number of disparate aspects of human cognition into integrated models. But integrated model cognitive architectures currently lack important capabilities, such as commonsense reasoning and higher-level cognition involved in speech and thought. Nevertheless, these systems—and the cognitive architecture project more generally—are interesting contributions towards developing more general cognitive AI systems. They differ from tasks in other sub-fields in being more ambitious in requiring and aiming towards more general overall performance rather than being at least in part, the engineering of systems that work well on specified tasks (G. Marcus et al., 2019)

27. Conclusion

A truly human-level, general Cognitive Artificial Intelligence (CAI), must be able to model human cognitive symbolism: the structure, origin, and situation of the meaning of everything humans care about, taught to them, very mostly inductively, during the years of their childhood, by their surrounding example and natural language. This is a generalization, specifically applied to modeling the part of meaning which is related to very low-level interaction upon and with the world and the social humans are interacting with and evolving into it as they grow up: the most and very human-like part of CAI's motivations, sentiments, social emotions, and knowledge. Meaning that, at the center of the CAI symbolization and cognition model, we must find the so-called "primary meanings": the multi-dimensional and multi-modal conceptual symbols upon and with which humans attribute and connect meaning to everything they care about: referents, the situated conceptual content or internal situation model of each referent. Everything subjectively perceived as existent by humans, in reality or in imagination. Symbols that encompass and are able to reflect a representational model of whatever humans learn or interpersonal social data which existables upon and within. These primary meanings are composites of the most elemental form of meaning: the assessment of orientation intent

as sensed by any subject within. Meaning that, if CAI wants to consummate on its own and independently human-like sentiments and the inherently human-like attitude toward Human Self as Made of Symbolic Minds, a principle and concept that defines Psychological Anthropology: the must-have social modeling enablement by which humans evolve into socially interacting individuals; must refer upon referential social symbols shared with humans, and attached to their derived spoken natural language.

References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? Oxford University Press.

Boden, M. A. (2016). AI: Its nature and future. Oxford University Press.

Cath, C. (2023). Artificial intelligence and ethics. Polity Press.

Daugherty, P. R., & Wilson, H. J. (2018). Human + machine: Reimagining work in the age of AI. Harvard Business Review Press.

Dehaene, S. (2020). How we learn: Why brains learn better than any machine... for now. Penguin Random House.

Eisenstein, J. (2022). Natural language processing. MIT Press.

Frankish, K., & Ramsey, W. M. (Eds.). (2023). The Cambridge handbook of computational cognitive sciences. Cambridge University Press.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2018). Cognitive neuroscience: The biology of the mind (5th ed.). W. W. Norton & Co.

Grossberg, S. (2021). Conscious mind, resonant brain: How each brain makes a mind. Oxford University Press.

Herrmann, D. J. L., & Frankland, M. J. (2022). Memory systems in artificial intelligence: Bridging cognitive science and machine learning. MIT Press.

Hitzler, P., Krötzsch, M., & Rudolph, S. (2022). Foundations of semantic web technologies (2nd ed.). CRC Press.

Horvitz, E., & Russell, S. (2022). The future of AI: Opportunities and challenges. Cambridge University Press.

Inaba, M., & Inoue, H. (2023). Cognitive robotics: Concepts and applications (2nd ed.). MIT Press.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

Langley, P. (2022). Cognitive architectures. MIT Press.

Linden, D. J. (2023). The accidental mind: How brain evolution has given us love, memory, dreams, and God. Belknap Press of Harvard University Press.

Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Knopf.

Murphy, K. P. (2022). Probabilistic machine learning: An introduction. MIT Press.

Panda, S. P. (2025). Artificial intelligence across borders: Transforming industries through intelligent innovation. Deep Science Publishing.

Parsons, T. D. (2021). Affective computing and artificial intelligence: Tools, advances and implications. Cambridge University Press.
Russell, S., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson.
Sejnowski, T. J. (2018). The deep learning revolution. MIT Press.
Shneiderman, B. (2022). Human-centered AI: Designing for the age of artificial intelligence. Oxford University Press.
Szeliski, R. (2022). Computer vision: Algorithms and applications (2nd ed.). Springer.
Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2022). Sensing as a service for internet of things: Big data analytics and cognitive computing. Springer.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-611-9

DeepScience
Open Access Books

Chapter 2: Building Blocks of Cognitive Architectures

1. Neurosymbolic Systems: Merging Symbolic Logic with Neural Nets

1.1. Introduction to Neurosymbolic Systems

Neurosymbolic systems are hybrid systems composed of both neural symbolic and purely symbolic AI capabilities merged together, delivering what both sides promised individually. Their goal is to bridge the two perspectives on cognition represented by connectionism in AI, the vision of artificial neural networks as models of neural computation capable of achieving general purpose intelligence, and classical symbolic or high symbolic AI, cognitive theories aimed at completely specifying systems of symbols and rules allowing the step-wise execution of reasoning procedures for solving complicated cognitive tasks. We will discuss some aspects of these two cognitive architectures in the following two sections, but our focus is chiefly on the connection-based models, and particularly, on how they relate to the symbolic aspects. You just have to gather that purely neural networks cannot do without symbolic specifications of what they apply to, and that purely symbolic systems cannot do without connectionism for the same reason.

1.2. The Role of Symbolic Logic

Let us briefly summarize the role of symbolic logic in cognition modeling, and more particularly, in knowledge representation and reasoning. Consider a cognitive architecture designed for commonsense reasoning tasks. Involved is a knowledge base containing a very large store of domain knowledge —built by some procedure, enumerating plausible reasoning patterns, such as for example a person entering a cafe and ordering for a cappuccino the following sequence of events: arriving at the cafe, ordering coffee, ordering the right kind of coffee, paying for it, promote some business goal of the cafe—and at least one inference engine capable of acting upon the stored knowledge, solving, or at least, producing plausible solutions for commonsense reasoning problems based on it.

1.1. Introduction to Neurosymbolic Systems

What can be more joyful for an individual than exploring their surroundings - be it nature or artificial, a social life or a system of knowledge, blocks of steel or blocks of numbers? Such exploration eventually leads to cognition, encoding knowledge, and eventually wisdom. We aim at building a machine that would be able to replace a baby for the idle of play, observe, supervise the process of existence, and eventually become wise itself. While there are systems that mimic human and animal cognitive characteristics, none of such have been able to reach the end goal of being yet another conscious partner for existing in what we call reality.

We are doling the building blocks out of symbols and logical operations, blocks of symbolic cognitive systems. These systems use logic as their main operational process for communicating with each other and the external physical world, as for also recalling what was acquired using perception and language. Based on a long period of research in different fields of cognitive robotics, artificial intelligence, linguistics, cognitive psychology, and early development of children, we present a neuro-symbolic architecture that is intended to realize communication, perception, and symbolic acquisition of knowledge within a hybrid logic-neural operator motherboard. Cognitive robots are machines that simulate the ability of human and/or animal beings to satisfactorily complete a variety of tasks that have a cognitive nature. Such tasks include the acquisition, use, construction, analysis, and recall of information and knowledge that were modeled or represented in some appropriate format for such – language and symbols, visual, spatial, and sensor-motor, colors and textures, sounds and sound filters, as well as information about the agent itself, its past experience and modeling of the environment and of social life (B. Hammer et al.,2023)

1.2. The Role of Symbolic Logic

Let's begin with the symbolic logic underpinnings. There are general observations about what an expressive symbolic logic includes or has, that we want our symbolic understanding module in our system to represent. First of all, there should be a generalized quantifier, giving us the ability to represent statements ranging from specific statements to generic or even indefinite ones. Secondly, we want the ability to represent conjunctions and disjunctions, statements that combine multiple statements together, either definitional statements about how combination values are related, or else disjunctive statements stipulating when a statement is supposed to be true. Thirdly, we want statements to be available in more than simply a propositional form, i.e. there should be first-order predicates at the basis of a natural language, that express relations amongst arguments, as there is no intent for the base model in our longer plan to destroy all rich representations and convert everything to propositional forms, making neural nets the only way to express relations rather than symbolic logic. Moreover, defining properties of quantifying expressions, that would need to be bullets on their own, can constrain the form of terms in that there appears to be something to skin covering the topic of weak constraints on statement forms of different logics (L. C. Paulson, 2020).

Technical points such as these are designed not to be obstacles, and there are techniques that make reasoning with second-order quantifiers, higher-order predicate logics, modal logics, equality predicates in logics with unique whereas inequationality predicates would lie behind a barrier that is related to the expressiveness of further building blocks, requiring more ambitious then message passing or simple supervised models to be surmounted. However, at the stage beyond language where a model has learnt to predict statement truth values, then pseudo-metrics will require calibrated relations that expand beyond the architecture basis, as has been observed where embeddings in models closer to lower-dimensional spaces shall not encode training examples at the stage of use.

1.3. Integrating Neural Networks

Feature Extraction

One of the first uses of neural networks in the context of cognitive systems is to extract features from some low-level input (for example images or sounds) and create a semantically richer symbolic representation. From the perspective of neuroscience this is related to the operations in the ventral visual pathway from the LGN to IT in the model of vision in which the main problems are to extract spatial frequency features and figure-ground separation before evolving to increasing levels of abstraction and object specific representation in the visual system. The made problem is computationally difficult and artificial neural networks (specifically, convolutional networks) have been shown to be particularly effective. The task of interest is to classify a visual scene as, for example, one of 1000 classes (F. Chollet., 2021)

Neurosymbolic Integration

One of the earliest examples of classifying an input with a learned feature set that uses symbolic logic is that of the neuro-symbolic object recognition system involving two parts: a CNN to produce prototypes for the individual image segments and a logic engine to reason about how the prototypes contributed to a solution (or, more commonly what the CNN had labeled the prototypes). In their work, the CNN was unmodified except for a softmax layer in the final section and they were able to include the CNN in the forward operations (a form of feedback). The reasoning about the symbolic components was accomplished by a first-order logic program which guided the recognition process and produced explicit labels for the semantic objects in the image (R. Besold et al., 2023)

1.4. Applications and Case Studies

Over the years, there has been growing interest in the potential for cognitive architectures to address artificial general intelligence. However, the most comprehensive example of CNSs is found in the human intellect. While machine learning has delivered state-of-the-art performance in visual perception, games, and NLP, there is still no solid theory on how pieces fit that could ultimately lead to the sort of general intelligence found in primates, let alone humans. Rather, the majority of successes have taken the form of spectacular breathless demonstrations focusing on individual tasks such as abstract painting, defeating grand masters in board games, or chatting like a novelist. Indeed,

those tasks can hardly be considered connecting or linking them, as claimed by Marr. There is no one-size-fits-all formula detailing how we can obtain general intelligence. Caveats aside, neural-symbolic systems — those for which some sort of NSyS integration is a key factor in their operation or sought-after attribute — have proven useful in many ways, and there is an expanding number of application areas. The primary object of the following sections is to enumerate advanced NSyS systems and their application areas. However, rather than a mere catalogue to be consulted when browsing the enormous literature on the topic in question, it is characterized by a careful selection of the most important models, architectures, systems, and technologies describing the vast NSyS landscape over the last four decades as a guide for novices and researchers unfamiliar with the area. The second purpose of these sections is not only to describe the basic ingredients and the various ways they can be combined, but also to attempt to articulate principles of design so that the right techniques can be applied to specific tasks and applications domains. This makes us believe that although according to the definition given above, NSyS cover a narrow aspect of the general or global functioning of CAs, they nonetheless remain an extremely relevant aspect that deserves to be according to its legacies and its strengths (P. Langley, 2022; G. Marcus et al., 2020).

2. Knowledge Graphs and Memory-Augmented Networks

Any cognitive architecture must be able to acquire knowledge and store it in an appropriate structure, a knowledge representation so to speak. In Artificial Intelligence, the predominant knowledge representation framework has been that of symbolic representations, in the form of rules and logical propositions, and the problem of creating a rich store of such representations has been tackled through the use of ontologies and knowledge graphs. In accordance with the AI revolution of the last decade, and particularly the idea of deep learning, cognitive architectures have supplemented symbolic knowledge representation underpinnings with richer sources of knowledge to draw from. In particular, retraining large neural networks on ever larger data sets has shown that it is possible to also create very rich sets of knowledge representations in the weight values of such networks. The memories thus created through training of such models have been dubbed both "knowledge graphs" and "knowledge bases" by different researchers in the area.

For deep learning to approximate any cognitive process in such a manner as to inform about the underlying cognitive architecture, deep learning must not only create semantic representations, more specifically knowledge graphs, but those semantic representations must also reach out to the inputs that flow through the network and modulate its processing at all levels of abstraction by the process of neural attention. Memory-augmented neural networks, such as Recurrent Neural Networks or Transformers, can be implemented with specific graph-based neural attention but attention is not necessary to be found at all levels, nor modulate, necessarily smooth, memory access.

2.1. Understanding Knowledge Graphs

Representational knowledge is essential to cognition and intelligence, and for a system to exhibit more general forms of higher cognitive functions, such as reasoning and learning, memory becomes an increasingly important aspect. Knowledge graphs have received significant attention in the AI community through their favorable use in supporting such capabilities, and they are now a vital component in multiple domain applications. Knowledge graphs represent concepts as nodes and relationships between nodes as edges. Concepts can be both concrete and abstract. Knowledge graphs may be directed or undirected, and they can have different numbers of edges connecting each pair of nodes. Knowledge graphs can also be composed of different types of nodes and edges representing different modalities of relations or concepts. Knowledge graphs can also include meta-related facts that describe properties of specific concepts and structured events involving specific concrete and abstract concepts. Multi-modal knowledge graphs can be composed of nodes of different modalities such as images of specific concrete objects or sounds of specific abstract concepts (F. Gandon et al., 2023)

The area of knowledge graphs has seen exciting advancements during the past decade in areas like knowledge construction, knowledge population and completion, and knowledge graph reasoning. Recently, employing deep networks and neural networks, in particular, to address these known challenges has received very significant attention. The area of knowledge graphs has also seen recent focus around foundation models applying transfer learning from mass data over weeks to months. Interest in learning from knowledge graphs using supervised, semi-supervised, and unsupervised learning has also received much focus, with related latent variable models being proposed to learn from partial or incomplete knowledge graphs.

2.2. Memory-Augmented Neural Networks

LSTMs, the predecessor to the Transformer architecture, were long thought to be the key progress in Deep Learning to solve the problem of vanishing gradient flow. However, MANNs came after, receiving far less attention in practice. The premise for these models is to reintroduce the idea of having memory explicitly defined during the model's forward pass and this memory being augmented during training, regardless of task. Similar to knowledge graphs, the architecture has a model-internal structure that allows for an attention mechanism to interact not only with input and output, but also modulates which memory to read and write richer than base Transformers (H. Jaegle et al., 2023)

These models can be seen as a bridging step between Neural Networks, where the vocabulary is likely to keep being relevant only for Routing Networks and Graph Neural Networks, whose memory inflation during the model's forward pass allows for detail to be ignored completely, focusing only on the most salient information from specific information sources. In hindsight, knowledge graphs are a generalized form of MANNs, where the memory is always visible during the model's forward pass, while MANNs are BERT-like architectures that generalize Transformers with external memory.

MANNs have a few possible training strategies, such as memory augmentation during supervised task finetuning, self-distillation from a teacher model using Inverse Denoising Autoencoders as joint training objective, reinforcement learning, unsupervised pretraining for retrieval and question answering tasks and embedding learning from data. MANN pretraining generally outperforms heavier models such as BERT.

2.3. Use Cases in AI

The incorporation of external memory into neural networks models in AI has opened up exciting new avenues for research in both modeling and application domains. In terms of application domains, the most common use case has been use of an external memory module to assist with natural language understanding, where the knowledge integrated into the memory is drawn from an established resource. Memory networks have been used to preprocess images and to guide generation of neural images; they have also been embedded in dialogue systems to allow them to learn about users on the fly. However, LM-based use cases currently remain limited almost exclusively to language models and dialogue systems. The ability of LM models to refactor their internal structures as they are pre-trained and fine-tuned presents amazing possibilities. In particular, this adaptable memory is especially useful for tasks where few examples are present during fine-tuning (S. Roller et al., 2024)

It is also now well-known that LM-based methods can produce state-of-the-art performance on a variety of popular NLP tasks – therefore, even though they may not be incorporated with the deep learning model of choice, language models with adaptable memory structures have directly influenced our ability to do good NLP work. Considering the variety of memory types that have been incorporated into deep learning techniques in their relatively short history, it will be interesting to see what the next few years hold. Populating the memory space with external knowledge and dynamically using this to influence model parameterization for specific tasks or domains provides a new avenue for research in both modeling and application domains.

2.4. Challenges and Future Directions

Though the field of data management for cognitive architectures is still at its infancy, the surge in recent years of work on memory-augmented architectures in the computational neuroscience and Artificial General Intelligence communities suggest that a wealth of opportunity lie ahead. We believe that several connections may yet exist between traditional ideas of semantic and episodic memory and tools like knowledge graphs and memory-augmented neural networks. First, though growing less common, the cognitive architecture community studies symbolic reasoning with some its goals involving replicating human behavior in tasks that require such reasoning. The construction of knowledge bases -- typically performed by human developers or workers via explicit programming or outsourcing actualization tasks to content developers -- has long been thought as prone to errors. Furthermore, the completeness of knowledge bases has long been questioned, often leading to the inability to fetch the right data when such

repositories are used. More granular ideas like ontological networks that allow knowledge acquisition via language grounding in addition to human labeling, that is transferring the symbolic to the sub-symbolic retrieval and organization of information in the network ledgers, have been proposed as a way to improve the robustness of core cognitive processes that require the use of semantic memory. In addition, recent advances in deep learning have opened the possibility to self-programmed architecture configurations based on unusual uses or likeliness of event sequences and for co-learning where symbol and sub-symbol systems help each other to overcome their respective inadequacies. Finally, when one studies typical trajectories of human development, its clear that expert knowledge cannot be simply shoved at humans via a curriculum of pre-existing data points shared across individuals and development hyper-parameters (M. Nickel at al., 2024)

3. Attention Mechanisms

Cognitive processes do not operate uniformly across the graphical landscape of cognitive architectures. It has become increasingly clear that selective processing of relevant environmental structures and, correspondingly, active inhibition of irrelevant structures, is one of the highest level organizing principles of cognition. Aspects of this theme have been applied in the context of memory operations, the flow of processing within architectures through great story brands, and the automatic, often unconscious regulation of structure activation strengths. However, these previous treatments have only tangentially addressed one of the major components of cognitive theory: selective attention. Selective attention is a process by which certain information is chosen for further processing while other information is discarded. Due to the limited capacity of human cognition, we cannot fully process all things in our environment; hence, a certain amount of information must be disregarded.

When highlighting the action known as focus, selective attention provides a means to allow certain inputs, memories, or task components to be processed above others. It allows input to be prioritized and focused upon when necessary. It allows information that is currently being used as well as task-relevant information to receive more cognitive resources than their current processing priorities would indicate. In this manner, it can be an aid to the selection, specification, and organization of outputs, modulating response characteristics. Thus, it fulfills both the high level and low level requirements of attention processing.

3.1. Introduction to Attention Mechanisms

The concept of attention spans a wide range of use and meaning. In human cognition, attention is typically a mechanism that selects a limited amount of information to process. As a descriptor, “attention” is also commonly used for shorter time projections, such as exposure, gaze or making a decision upon. Often human attention is modelled as a filter, integrated over space and/or time, which can be understood as a probability

density function over either of the domains. Additionally, the search for saliency maps – images in which the most salient features are emphasized – has been a topic of research in the cognitive and the technical domains. The modeling of attention in different modalities has supplied additional components for attentional models, such as modulation, reactivity, and memory ties.

Using this definition as our foundation, we can transfer it to the domain of artificial intelligence (and, specifically, neural networks). Artificial Attention Mechanisms are designed to assess how much "attention" – in the sense of the use of energy and/or processing resources – should be allocated to which part of the input at which point in time, and what, how much, and how quickly factual task-related changes in attention should induce. The result would either be modifications of the signal per se (such as amplifying certain features), or in the way how the following transformations are being weighted (for example, such as emphasizing certain patterns of the learned model). Accordingly, Attention Mechanisms can act either as multipliers of feature channels through space, or as modifiers of temporal processing flow. Processes governed in such a way naturally allow for the modeling of long-term dependencies (D. Bahdanau et al., 2024)

Attention Mechanisms have been independently discovered and applied in different areas of engineering, such as computer vision, natural language processing, reinforcement learning, robotics, and cognitive modeling. Much of the interdisciplinary exchange between cognitive science and technical development that has fueled the exploration of Attention Mechanisms stems from developments in Neural Networks and, in particular, Deep Learning. The recent interest in modular Neural Network architectures has re-fueled the investigation of Attention-like designs for image transformations.

3.2. Types of Attention Mechanisms

Attention mechanisms can be categorized depending on various criteria. One such criterion is whether or not the attention is defined in terms of the entire input sequence. In global attention, the attention weights corresponding to the output step t are computed using the hidden states corresponding to all the input steps. In local attention, on the other hand, we compute the attention weights corresponding to the output step t only using a subset of the hidden states of the input sequence, sometimes including only the hidden state of the input step that corresponds to the closest time to t.

Another common criterion for categorizing attention mechanisms is whether they perform hard selection or soft selection. In hard attention, which is rarely used, the attention mechanism selects a subset of the input hidden states. In soft attention, the attention mechanism assigns a weight to each input hidden state, where the weights sum to one, and the output context is a weighted mixture of the input hidden states. Due to its parallelizability and differentiability, soft attention is better suited for incorporation into standard neural network architectures and has consequently been used more frequently than hard attention (D. Bahdanau et al., 2024)

Additionally, from an implementation point of view, attention mechanisms can also be categorized into two types depending on how the context vector is computed. In additive attention, the context vector for a particular output step is computed as an intermediate, whereas in multiplicative attention, the context vector for a particular output step is computed in a single step. The term memory is used in some definitions of attention mechanisms to refer to the set of input hidden states from which the context vector is calculated.

3.3. Impact on Neural Network Performance

To understand the impact of the attention mechanism on the performance of neural networks, it is important to understand some details of the attention models and of the benchmark tasks that came to use attention models. The transformer was introduced as a general-purpose architecture for Natural Language Processing tasks, and was shown to set new state-of-the-art performance levels in previous pre-trained recurrent language models. Later research explored a hybrid model configuration, where pre-trained CNNs or RNNs are further fine-tuned on downstream tasks using the transformer. Few-shot in-context learning with language models was introduced as an approach where feature extraction is performed with the language model itself, using the examples from the task description as input, where the input is concatenated to task description. Note that in-context learning has been shown to work much more efficiently with larger language models and with few examples, compared with fine-tuning approaches (J. Lin et al., 2023)

A key idea in recent large language model design is scale – the use of billions of parameters and training on datasets containing an order of magnitude more data than used in previous state-of-the-art models, which were trained on hundreds of gigabytes. A surprising aspect of the scaling of pre-trained language models is the results on few-shot in-context learning. Recurrent language models, which use a similar architecture and similar scale pre-training of additional objectives and smaller scale datasets and are fine-tuned on downstream tasks, do not show the same in-context learning capability. It has been hypothesized that in-context learning is made possible by the attention mechanism – the computational arrangement between input and output conditioning. While not confirmed, there are indications that using efficient computer hardware for training the large transformer models and for their inference is a key factor in their success.

3.4. Real-World Applications

As described in Section 3.3, attention mechanisms have been shown to have a profound effect and positive impact on a neural network's practical performance and efficiency. Deep neural networks for a variety of different applications, such as video understanding, computer vision, document understanding, speech processing, and music processing, are all enhanced through different forms of attention mechanisms. Attention mechanisms within the computer vision domain generally allow neural networks to interrogate image pixels or regions, understanding which pixels are salient and

contribute most to the output of the neural network. This is useful for tasks such as detecting and identifying objects in images and videos, which can then be utilized in other programs to process and manipulate the visual data. Camera-based object recognition can be sped up using active computer vision models that make use of attention mechanisms to quickly determine key image regions. Attention mechanisms are also widely used in machine translation tasks. Neural network models that encode a source phrase into a hidden representation and create a target phrase word by word are enhanced using attention mechanisms. Each of the individual words in the hidden representation, which are linked to specific source phrase words, can be focused on at different points in the model's decoding process, enabling the neural network to closely mirror how human beings translate phrases from one language to another. A critical requirement of machine translation systems is that they be real-time and fast; the wide range of different machine translation models that utilize attention mechanisms are able to translate phrases of varying styles and complexities faster than traditional translation programs (J. Lin et al., 2023)

4. Transformers and Long-Term Memory

Transformers are deep learning architectures that use self-attention both to implicitly compress data and to compute an explicit type of long-term semantic memory of relative meaning. They have been shown to be excellent building blocks for natural language processing. However, they have yet to be shown to be effective as cognitive architecture modules beyond language. In this section, we explore the memory capabilities of Transformers, with an eye toward future prospects both for extending their capabilities and for repurposing them in other ways as cognitive architecture modules.

Overview of Transformer Architecture

Since their introduction about five years ago, Transformers have become the new gold standard for natural language processing tasks. What sets them apart is twofold. First, they use multi-head self-attention on digital data. The self-attention computes a complete semantic memory of all elements of the input, which is then used to condition the meaning of other elements in the input. In their encoder-only version, BERT and its successors are currently considered the best models for natural language understanding tasks. In their decoder-only form, GPT and its successors currently dominate natural language generation. In their encoder:decoder form, T5 and its successors have had successes in both directions, both NLU and NLG. The multi-head self-attention allows for more complex interactions among the latent dimensions, at a small cost, in terms of memory and compute, compared to the single-head attention mechanism in earlier LSTMs and GRUs.

4.1. Overview of Transformer Architecture

In recent years, the transformers have radically transformed the field of deep AI architectures. Thanks to super-computers trained on massive data resources, these neural network architectures have significantly improved performance in many problem domains like natural language processing and image processing. We now describe the main building blocks of a transformer architecture. We begin with an overview of the architecture, followed by details of the modules used in one established implementation. A model consists of a stack of identical layers containing two main modules: a self-attention module and a feed-forward neural network module. Both modules are surrounded by layer normalization and residual connections. The transformer architecture was inspired by the use of attention by neural machine translation models and recurrent neural networks, which leverages context words to compute a representation of the current word in the vocabulary that is semantically related to context words. In turn, the transformer architecture led to further investigations of the attention concept, such as the introduction of varying other modules of transformer-based architectures by speeding up the training process with knowledge distillation, by enabling the processing of arbitrary-long input sequence tokens, and by adapting the architecture for use with time-series and tabular data (A. Vaswani et al., 2023)

4.2. Long-Term Memory in AI

Long-Term Memory (LTM) serves a crucial and unique role in cognitive architectures, setting them apart from more specialized systems. In the context of transformers, LTM specifically refers to the means by which knowledge is preserved in AI systems over long periods of time. The importance of LTM in human cognition is evidenced by the manner in which it is packed with information through a lengthy and demanding process of education. Cognitive architectures take this further by emphasizing the need for an LTM that supports extremely diverse, domain-general cognition. In CAs, LTM addresses the functions typically carried out by the natural-language-processing pipelines: serving not just to store information, but to make predictions, draw inferences, and power other kinds of reasoning.

Ideally, LTM systems would achieve this general mental function by containing a vast resource of native knowledge represented in a special, specialized format. Yet engineering such a large-scale system is an undertaking notoriously beyond human capabilities. It is not yet clear whether or not transformer-style LTM research will yield systems that support cognitive functionalities, especially for systems that refer frequently to external information. The more ordinary domains of LTM research into knowledge distillation and knowledge transfer assist transformers in maintaining static information or imitating the behavior of yet-to-be-trained neural networks or never-multiple-task training scenario. Researchers are currently looking for ways of extending LTM functions to cases beyond distribution shift in hopes of matching CAs in their influential capacity for domain-general activity (I. Goodfellow et al., 2022)

4.3. Implications for Natural Language Processing

The preceding sections have reviewed the mechanisms of transformer architecture and its performance uncertainty. Detecting and generating potential correlations in variable-length data is a fundamental process of human intelligent activity. The performance of transformer and language model research in NLP naturally leads to a question: Does the architectural design of a transformer model match the way humans develop cognition based on linguistic communication?

This question prompts us to discuss the potential developmental implications of cognitive architectural theory. Neural communication has become an important pathway for exploration in deciphering human brain functions. As a special form of neural communication, language plays a unique role in neurocognitive sciences. During neuroimaging, neural networks generate synchronous responses and engage in information exchange. Some researchers have proposed the concept of neural language of the brain. During neuroplasticity, concentration-modulated learning and synaptic strength modification are key mechanisms for the change of spontaneous neural communication and brain functions in neurodevelopmental process. In addition, neurodiversity emphasizes the cooperative activity of multiple brain areas within the framework of neural language for normal functioning in human communication, which is supported by findings that cognition involves communication and coordination between multiple brain networks. The neural communication criteria require language-like features of neural communication; that is, the communication signal should exhibit compositionality, discreteness, and recursion. Neurocognitive science research results reveal that language is an important psychological process that constructs thoughts and memorializes object properties and relations; however, the theoretical foundation supporting the design of neural language, including approaches employed in NLP, is still unclear (J. Devlin et al., 2022).

4.4. Future Prospects of Transformers

Promising avenues of exploration regarding Transformers include accounting for pretraining self-supervisory learning across multimodal modalities and semistructured data; accounting for learning inductive biases across training data distributions; scaling algorithmic propositions for application to inputs; estimating latent hidden states with neural attention; neural autoregressors with long-range dependencies equipped with long-range multiheads; and bridging statistics to computation. Addressable learnt latent hidden states can be optimized algorithmically, ultimately computing with statistical inference methods eigenparameterizing a memory matrix such as with large language models for pattern recognition-based information retrieval on tasks such as question answering.

Learning distribuitional representations can increase the training locality of recognition memory typically implemented through a memory matrix for supervised learning, so that the memory requirement scales logarithmically or polynomially with the sample space size. Implicit manifold representation can further achieve independence of the sample

space size with access to an activation function with exponentially increasing slope, such as the rectified linear units. Copyright explicit million-sample naive Bayes memory for a million-sample English-structured language-based set, and deepen captured patterns with residual structures. Elucidating the desired range of the activation function can additionally yield representations to minimize optimization time, leading to effective million-sample implementations.

Neural Networks have been collecting accolades in recent years for their extraordinary performance on complex high-dimensional problems in text, vision, and time-series across diverse real-life tasks. Transformers have been particularly celebrated due to their flexibility in handling sequence-like inputs/outputs and the availability of transfer learning. Both theoretical and empirical results have underlined the tremendous capabilities of Transformers beyond just brute performance (I. Goodfellow et al., 2022).

5. Reasoning Engines and Logic Inference in Modern AI

With the surge in recent years of data-driven AI systems, it is important to remember that the aim of intelligent systems is to emulate the capabilities of the human being, and that logic reasoning is a key aspect of the high-level intelligent processes in humans. From purely biological aspects, all human beings undergo the sort of logic reasoning processes proposed by the psychological development of these processes. If we try to understand how a child gains and develops his/her ability to make uses of reasoning systems and logical inferences, we can compare his/her progress with that in AI systems. The child, from the incredible amount of stimuli received, creates different beliefs about the world. These beliefs serve as premise data for later reasoning processes, when the logical inference mechanisms are applied to these beliefs to create new beliefs from the old ones stored in memory. These new beliefs enriched the knowledge structure in a pipeline-like mechanism. This is the particular system of logical inferences and logical reasoning that this study is addressed to. It is more the reasoning level and functions than the contents of human knowledge representations that are of interest.

There exist many logic inference approaches and techniques, and there have been a remarkable amount of research dealing with these topics. We can find any logic application for making a known logic inference, within different AI fields. Reasoning engines appear in a broad range of AI works, from simple applications performing only a single inference, to completely developed functions present in general AI problem solvers. However, the majority of reasoning engines are not exclusive for AI. Most of them are not originally from AI. Many of the well-known existing logic inference techniques were developed in the area of mathematics and computer science. For that reason, logic inference can be referred sometimes as an “off-the-shelf” technique that can be used within AI systems. It is important to keep in mind that there are many logic techniques in the researchers’ hands and that some of them are not in AI works or designs, not even have a relation with the AI systems. Notwithstanding, no doubt that

such techniques are the basis of the developed AI applications. This is the support of intelligent function.

5.1. Understanding Reasoning Engines

After many years, cognitive science is now facing important advances, in some specific problems of Natural Language Processing, Image Recognition, Knowledge base creation, and automated problem-solving. These are elements of intelligence, aspects of the global model of cognitive systems. A decadent aspect of those areas, Logic Inference, is now, simultaneously with methods such as deep learning, receiving a new wave of boosted research in a very fruitful way. But not only poor and less essential tasks involving such logic knowledge representations. A hierarchy of cognitive systems is being created to have specialized systems doing the visual perception, the dialogue, the intrinsic knowledge representation, and the knowledge inference of the collaborative and decision-making processes of the global systems.

The new type of engines that augment classical search engines, Inference Engines, are now being created to sustain or cooperate with all the other AI processing components. This enables, associated with the Data with Knowledge interfacing, a new generation of cognitive systems that have the global performance needed. This paper discusses this area of knowledge, but with a total cognitive system's perspective, going from the perception to the motor coordination, but concerning mainly Knowledge Inference, which is its strong area. After this introduction, we briefly do a survey of the area, defining some specific terms needed, and how it integrates into the new generation of cognitive systems. After this, we present how we can nowadays exploit Logic Engines, searching where and how such techniques can help in the joint task of knowledge representation and inference. Then we discuss the existing logical programming languages, showing which logic problem-solving methods can be successfully used in each of the planned system modules. Finally, we summarize the paper stating its most relevant conclusions and future trends of knowledge inference on cognitive systems, highlighting the necessity of the collaboration of the logic community with other community's AI areas implicated (Goertzel, B et al., 2020).

5.2. Logic Inference Techniques

The logic methods can be classified along different lines of variation: the logic of inference, the knowledge representation with which it works and its underlying architecture. Among the many logics of inference, we distinguish propositional logic, many-valued logic, temporal logic, modal logics, dynamic logic, and sub-structural logics. Each of those logics can be implemented over different axiomatisations corresponding to optimized calculi, developed languages and/or semantics. The Knowledge Representation Language the logic operates over is usually a variant of predicate logic, ranging from the most general where horn clauses over predicate calculus are used to express logical forms, to restricted languages such as logic programs or modal logics variants. Different axiomatisations route different theoretical properties

or efficiently optimized systems, e.g. whether or not additions and/or reductions of logical forms are performed, or how they are implemented.

Most Cognitive Architectures deploy a two-valued propositional logic with a restricted Knowledge Representation Language they use to implement a Logic of Inference core, endowed with additional knowledge modules that extend its intellectual capabilities. The Reasoning Engines corresponding to those Architectures have been tested in laboratory two-parts implementation scenarios. These found large superiority margins of classical model-based BCIs over neural models, in the first testing case: performing temporal anomalies on counting activities of imagined technically simple hands movements. But they missed the more general allegations of the LIDA Architecture, along the 'wild charm' against traditional AI explained in the introductory chapter. Those efforts validated procedures of reasoning predictive behaviors, by means of Cognition Feedback and Predictive Memory. But further practical validation is required to transfer these models to long-term normal daily routine activities of a cognitive agent (M. Genesereth et al., 2021).

5.3. Applications in AI Systems

The analysis of the application of reasoners included in the design of intelligent systems shows convenience to adopt different, complementary techniques. Reasoners based on production rules are able to accomplish well defined tasks at a very high speed, which is a convenience to accelerate system reaction. These systems can correctly answer to questions, but they are incapable of more complex elaborations which make necessary the generation of new knowledge. Reasoners based on inference with first order knowledge can accomplish synthesis tasks when trying to explain the given input data. In such cases the need to generate new knowledge is important to the solution of the task and requires additional mechanisms able to optimize the possible applications of the inference techniques. The reciprocal reinforcement between the complementary techniques has shown validity in many systems which have been proposed. Typically, these architectures fuse the complementary truth maintained at different levels of sophistication offered by actual production based systems and logical formalism. They have included means to get information structured at the intensity level in order to generate modules able to produce intermediate expert structures and the dependence with strength relations with respect to the overall system which have to be taken into account by the reasoners. The techniques to store relations with strength of the expert systems domain are positive localized advisors able to accumulate knowledge. By direct execution of the knowledge stored by FOL with respect to some actuality variable, the experts help logical inference establishing expert basis points or negative points in the case of low intensity signal (M. Genesereth et al., 2021).

5.4. Comparative Analysis with Traditional Systems

In this section we present a comparative analysis of some specific attributes of traditional systems against the considered high-level cognitive architecture. Potency and resource consumption establish a comparison regarding the use of logical inferences and the

conventional structure of traditional systems. Reliability proposes a different hypothesis on the broad definition of intelligence as states of a cognitive agent. Robustness turns the prism to the other side and studies in which conditions symbolical representations struggle to offer intelligent performances. Then, we handle the possibility of interpretation of an intelligent performance: transparence both for humans and for the AI itself.

As we said in the previous sections, the presented knowledge structure combines symbolic and connectionist representations. A reason that justifies this hybrid organization is that both high-level types of knowledge are used in different situations. Potency and resource consumption: It is unquestionable that highly structured knowledge avoids many high-cost repetitive or costly inferences. The typical complexity analysis of a semantic network doesn't provide a clear idea though. For instance, any network of mono-synaptic links will allow for efficient pattern matching. The semantic form of a complete mesh of ontology at any semantic depth would force a code of type O(max|n|.)-even for very simple inferences like a simple question answering where the answer is known to be in the knowledge base (J. Andreas et al., 2023)

6. Integrating Causality and Common Sense in AI Models

There is a common belief that while it is easy to build predictive models, these models simply learn to predict correlations in data, and this type of model is unsuitable for general intelligence. In contrast, symbolic AI is often thought of as utilizing common sense, domain knowledge or world knowledge. This type of knowledge is easily expressed symbolically, in languages like pierwszy iwa or Logic. Such knowledge enables a model to be efficient in formulating plans, making inferences about the world through reasoning, understanding and telling stories correctly, generating complex actions and should thus be utilized in order to build more clever models of AI. Causal reasoning is often considered as critical to intelligence, since without it, there is often a lack of understanding of the problem domain.

Unique to all intelligent agents is their ability to acquire, recall, and utilize common sense knowledge to succeed in all manner of artificial intelligence tasks. Agents not only acquire and utilize commonsense knowledge, they also expand their store of knowledge through reading and listening. Unfortunately, most intelligent agents today do not actively acquire commonsense knowledge, employing instead some version of deep learning trained on large datasets in order to perform their assigned tasks. These models learn correlations reliably present in those datasets, but ultimately fall short of achieving true intelligence because they cannot actively expand their commonsense knowledge, the information present in those datasets may be faulty, and are incapable of any kind of causally driven planning or reasoning.

6.1. The Importance of Causality

While we often tend to take causality for granted, recognizing it as essential for both human and animal understanding and reasoning, it is in fact a profound and tricky issue. Recognition of the primacy of causality in semantics is ancient, going back at least to Aristotle. A central argument in favor of the special or even exclusive importance of causal thinking has been the enormous effort in the last few decades to develop methods for inferring causal structure from data, often with statistical concepts based on conditional independence and use of sufficient conditions for such independence facts. But the sensitivity of various joint distributions to non-causal relations and the difficulty of comparing causal and probabilistic projectivity hypotheses remain fundamental obstacles. Moreover, the claim is challenged by philosophers proposing a projectivity alternative.

But the main goal of this introduction has not been to survey or justify the rationale for an architectural role for causality in AI, nor even to argue that absent such incorporation AI may be enhanced but not truly comparable to humans or, at least, the systems that they or other intelligent animals are. Rather, this introduction is concerned with clarifying our conception of that architectural role; in short, how should an AI architecture incorporate the causal concepts and processes? The relevance of this clarification makes it imperative to take account of the difficulties and issues raised by the alternative approaches to the relation between the projectivist, excess, and rationalist. In this piece of work, we sketch some answers to that question by specifying some desirable features that we think any appropriately integrated architecture should have (B. Scholkopf et al.,2022)

6.2. Common Sense Knowledge in AI

AI models are incredibly powerful, but their knowledge of the world is limited because they are often missing both important information, as well as cues provided directly by inference. This loss of cues would not be a problem if the model had learned enough of the appropriate knowledge, or commonsense knowledge, drawn from prior data, along with priors that allowed useful reasoning over this knowledge. For language models, their size — billions of parameters tuned for language — implies that they have access to enough “slots” for important bits of general knowledge. Classic back of the envelope reasoning, however, suggests that their access to commonsense knowledge is unfortunately quite poor. Before delving too deeply in that direction however, it is worth taking a quick detour to clarify our terminology regarding either classical or contemporary definitions of “commonsense” knowledge.

Commonsense Knowledge is knowledge that is common to most people, and which does not involve the particularities of a person’s background, or the current time or place. In common sense reasoning particularly, this knowledge serves as background reasoning which don’t want much detail to make an inference. Common sense is a necessary but not sufficient set of knowledge about the general facts about the cognitive, physical, and sensorial world of humans. It consists of basic facts about how people go to work every

day, and why they need to travel on the ground. These are facts which are not established by experts nor particular to a domain. These carry basic principles which should be learned or hold true for every conceivable person from any domain (Y. Choi et al., 2023).

6.3. Techniques for Integration

There are no existing cognitive architectures that yet integrate together causality reasoning and common sense knowledge representation, dissemination and reasoning. That said, there has been work on a number of models that are bridging some aspects of the two areas; so called integration works. We briefly describe a few of them here, emphasizing the techniques they use. Causality has both a functional role and an epistemological role. In the functional role, causality is used by the model to support a specific collection of functions, such as grounding and relational domains, or generating problem spaces or basic action hierarchies. In these models, there is no claim that the structure of the cause-effect infrastructure is an explicit foundation for the entire model – even though in believing-C that is an explicit assumption. Common sense is supported whole idea by the knowledge base: the model is not meant to behave as a general intelligence system, only to have a substantial pool of common sense knowledge.

In this paper, we have needed an important socle in common sense to help the classical ai systems become intelligent – and by extension the cognis-space as a whole. We show some techniques devices that allowed this integrated knowledge database. Other integration works on the AIXI model, specifically for the agent-AIXI. One of the main claims of this work is that, in order to generalize common sense knowledge, a small set of about one hundred axiomatic goals are sufficient. For knowledge compilation, works on deep neural networks realistically responding to the properties of common sense explanations. These works use Probabilistic Graphical Models to compile substantiated mini world of common sense knowledge in these DNNs (M. T. Keane et al., 2020)

6.4. Challenges and Opportunities

A number of challenges arise when integrating causality with common-sense knowledge. While learning from data, the former requires a structured search space that guides the process towards the identification of the right causal relations from a large pool of potential ones. DNN models are primarily data-hungry systems, and obtaining enough samples for alike events that can be used to elicit their causal relations can be difficult in complex systems that involve hidden variables. Explicit knowledge that can establish logical invariants that must be satisfied by all underlying causal models, the use of generative causal models for data composition and the boost of data through computational processes can be used to alleviate data-hungry learning. Common sense knowledge is both qualitative and quantitative, integrates facts and heuristics, and is often imprecise and fuzzy. On the other hand, learning such knowledge and representing it is a challenge in machine learning systems. In addition, the use of deep learning systems that can formulate general answers to general interactions but that do not provide any baseline explicability of their responses hinders its use for guiding AI systems to develop layer-wise automatic common sense reasoning steps. The human cognitive

system operates hierarchically, and motivations, emotions, social or common sense backgrounds and other aspects of lower-level reasoning must be topologically integrated in DNN architectures for allowing the automatic layering of common sense operations. Recent advances in AI, particularly the success of DNNs, often offer the opportunity to perform difficult tasks involving perception or reasoning by using pre-trained models. Large language models and multimodal models, that operate on inputs of different nature, have become widely popular and capable of impressive performance despite being blind to the topology of the mappings they implement. Such models are however fragile, justifying only some of their empirical successes, and integrating shape limitations a priori or a posteriori can help structure these large models for specific applications while generally boosting their performance (Y. Bengio, 2024).

7. Conclusion

The main idea behind this work is that if cognitive architectures are to advance in their development and experimental testing, they must do so in a complementary manner. Accomplishing this goal can be achieved through the specification of an architecture's most basic components and the assembling of existing cognitive facilities into modular, specified building blocks in order to form progressively more complex and specialized cognitive architecture configurations. These blocks can then become the means by which cognitive architecture experimentation is advanced by providing architecture models of increasing complexity and capabilities. Defining building blocks is an important and arguably necessary step toward modularizing cognitive architectures. Obtaining a more modular architecture system enables the architecture to evolve more rapidly, both in terms of theoretical accounts and applied systems, while at the same time providing computational building blocks that implement specific facets or capabilities of intelligent behavior. This, in turn, would then enable cognitive architectures to become consistency, multipurpose models of artificial and human intelligence.

Furthermore, cognitive building block definitions could then become a repository of computational functions that define capabilities of human and artificial intelligence. However, defining the building blocks for cognitive architectures should not be an isolated task. Such a process should take place with an eye toward bridging the gap between theory and application and informing the different communities of their similarities and differences. We hope that the definitions presented and their related categorizations will serve as the starting point and reference for such an enterprise and are subsequently enhanced through experimentation, results, and collaborative work. Finally, such a development of building blocks for modularization of an architecture system both in terms of ease of exploration and experimentation and actuation resource management would deliver the vision of architecture for lightbuilt robotic agents.

References

Andreas, J. (2023). Language, logic, and learning: Integrating symbolic and neural approaches. MIT Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2024). Neural attention mechanisms: Foundations and applications. MIT Press.

Bengio, Y. (2024). Deep learning for AI. MIT Press. (Forthcoming, early release online).

Besold, R., d'Avila Garcez, A., & Lamb, L. (2023). Neural-symbolic cognitive reasoning. Springer.

Brown, T., Raffel, C., & Vaswani, A. (2023). Advances in transformer architectures: From attention to memory. Addison-Wesley.

Choi, Y., & Sap, A. (Eds.). (2023). Commonsense machine learning. MIT Press.

Chollet, F. (2021). Deep learning with Python (2nd ed.). Manning.

Devlin, J., et al. (2022). BERT: Pre-training of deep bidirectional transformers for language understanding. Morgan & Claypool Publishers.

Gandon, F., Lefevre, F., & Zimmermann, A. (2023). The semantic web and knowledge graphs: Future directions. MIT Press.

Genesereth, M., & Nilsson, N. (2021). Logical foundations of artificial intelligence (2nd ed.). Morgan Kaufmann.

Goertzel, B., & Pennachin, C. (Eds.). (2020). Artificial general intelligence. Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2022). Deep learning. MIT Press.

Hammer, B., & Hitzler, P. (2023). Perspectives of neural-symbolic integration. Springer.

Jaegle, H., & Ha, D. (2023). Neural memory architectures: A deep learning perspective. Cambridge University Press.

Keane, M. T., & Smyth, B. (Eds.). (2020). Case-based reasoning: A textbook. Springer.

Langley, P. (2022). Cognitive architectures. MIT Press.

Lin, J., & Lee, D. D. (2023). Self-attention: Mechanisms and applications in deep learning. MIT Press.

Marcus, G., & Davis, E. (2020). Rebooting AI: Building artificial intelligence we can trust. Vintage.

Nickel, M., & Rosati, L. (2024). Representation learning for knowledge graphs. MIT Press.

Paulson, L. C. (2020). Logic and proof: An introduction. Cambridge University Press.

Roller, S., & Devlin, J. (2024). Memory-augmented language models and applications. Springer.

Scholkopf, B. (2022). Causality, learning and reasoning in AI. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. MIT Press.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-611-9

DeepScience
Open Access Books

Chapter 3: Toward Artificial General Intelligence (AGI)

1. What is AGI? Definitions, Debates, and Desiderata

In contrast to narrow AI, easy to already built systems that are useful yet short-sighted, yes potentially confusing agents that work awfully well according to simplistic but predictive results, AGI attempts to create systems that think as large cognitive apparatus, autonomously working long term to negotiate finding solutions to arrive at a trail of goals, for vague diverse teams of less well able humans. In this essay, I describe key themes and ideas I feel are central for building AGI.

The Autonomous Academic Agent, AAA, which started as a student assistant system throughout its near future but is lately getting people over its scale to self-study common text and try to solve problems while asking questions at human interaction windows, is acting towards developing AGI. Within their design, there are groundwork ideas for speculation in understanding the road that incrementally smooths a trail for coming large cognitive assistants, LC2As, and for future thinking colleagues and intellectual competitors, knowledge-economizing haters of huge amounts of page and lame image generations not called for asteroid or material resource pollution, which skew balance on numeric predictions.

But what is AGI? A focus on definitions would seem necessary to clarify what is intended by the expression. The acronym funneled attention to general forms of intelligence by its semantic denotation in a specific context. Generalization was put into the picture. The term was coined to enthuse hopes in a then small crowd of academic researchers exploring the creation of intelligences without mouths or legs or big arms outside our heads (R. Kurzweil, 2024).

2. Modularity and Transfer Learning Across Domains

Many aspects of learning and performance in animals and humans are modular in nature, and the question arises of whether this is also true in artificial systems. Some elements

of cognitive systems are easy to decompose into modules, including individual tasks, such as visual object recognition or game playing. What is less clear is whether approach, motivation, or emotion systems really do implement modules that have separable functions. A simpler form of modularity is that of functionally unified components that nevertheless interact heavily with one another and require access to information from all aspects of the system to compute their collaborated outcomes. For example, modularity and the idea of latent functions or conditions within motivational and emotional state or recipe systems are not mutually exclusive ideas. However, one normally hears the term modular to describe a functionally unified component that can function independently or is at least considerably insulated from other components or access to joint problem-solving is performed simultaneously or in a time multiplexed manner. These ideas of module interaction and integration suggest thoughts about transfer learning in the context of general cognitive architectures. It is commonplace to design specific systems to perform some well-defined task in isolation. However, humans readily transfer skills and concepts learned in other domains to new domains and almost always benefit from such experiences. Much work in artificial intelligence and robotics has been devoted to the opposite, incrementally building completely closed systems, which then must be tested and validated on whatever data happens to be available, without the use of domain knowledge. This may be possible for some applications, but it is hard to believe that this is the appropriate long-term approach to building general intelligent agents (Y. Bengio et al., 2023).

3. Multi-Modal and Embodied AI: Language, Vision, and Action

Tree of Thought reasoning has thus far been purely symbolic in that it has only dealt with tokens from the space of the static and propositional world model. Observing that agents in the world have in their toolkits not just languages for communication, but also vision for observing the world and action for changing it, there is also a natural question of how symbolic and non-symbolic AI could be combined. Multi-modal reasoning has naturally been at the center of embodied AI since language has always been one of its main components. From the earliest days of interactive language learning by children, parents and psychologists alike have recognized that children learn about concepts they hear in language by observing their parents using them in vision and action, often with their understanding grounded in a current shared visual focus of attention, such as a toy being pointed out by the parent. A parent can use words to direct a toddler while pointing, and then observe whether the child follows their instruction. Should the child fail to act as the word's meaning would suggest, the parent can correct or reconcile the toddler's understanding of "bird" as related to something other than the nearby visual bird image, such as the child's action of drawing close to the toy city below the towering toy block birdhouse.

Without the ability to act and see, words are empty of meaning, and some philosophers have asserted that language has meaning only when a speaker and listener share a current

common visual perspective or script. Other philosophers agreed that language is meaningless in isolation, but wanted to ground language not in perception but rather in the act of speech. Certainly, languages provide a most efficient way to communicate about visual perception and visual actions, and we have found this to be the case ourselves in our experiments with guiding robots to follow instructions (Y. Bengio et al., 2024).

4. Meta-Learning and Self-Improving Systems

Meta-learning, or learning to learn, is a promising avenue toward more generic and powerful machine learning behaviors, which may in turn create individual intelligences with greater ingenuity, depth, and scope. There are a number of aspects to this goal, including the phenomenon of few-shot generalization or the ability of many modern deep learning techniques to generalize from very few examples, but requiring a significant model investment to access the capability. In this work, we will primarily address the latter case, with more is better scaling guiding the process. The concept of few-shot generalization has also been discussed in the context of accelerated transfer learning and hyperparameter optimization, but there has not previously been a machine learning-level meta-learned model portfolio solution capable of transcribing across ambitious and diverse data modalities at human parity or exceeding measurable human capability with respect to scaling properties of model investment and distribution size.

Self-improvement is also an age-old concept of intelligence development and is evident across many different intelligence forms. Models, especially if meta-learned, should generally gain accuracy with greater scale if the flow of new or novel sample data improves over time. However, models seem to be limited to directional learning efficiency with each update not following an optimization problem and setting the next step in the correct direction. The apparent loss of steady optimization directions as models became largest for the tasks which they were intended to solve was indeed a concern and a known open problem which informally describes “safety”. Still, forward-creep update direction self-improvement velocity is highly variable, being long correlated with model sorting, portfolio construction, and training coupling while at times acting globally anti-cyclic as illustrated in the recent history of predictive model evolution. The mechanism and origination of bottlenecks, cycles, and other behavioral aspects of model evolution are only beginning to be explored (L. Zou, 2022)

5. The Role of Simulation and Imagination in Generalization

The precise character of generalized learning has long been a topic of great interest, since most of what people and animals learn is of this kind. Backpropagation neural networks learn primarily by an extremely simple rule. Despite their ability to learn an incredibly rich variety of tasks, and a plethora of practical applications, these tasks generally

express fully generalizable functions, or at least something which can be framed as such, such as supervised tasks on large, clean, referred datasets. As stated earlier, research has shown that tasks which bypass generalization entirely are much easier for backpropagation than the zero-generalization virtual data equivalent.

There is substantial evidence, applied to a variety of tasks across multiple domains, that animals and humans rely heavily on internal simulations and imaginings of the world, and that such imaginings underlie generalization. For example, the continuous or recursive use of simulated motion of objects through space has been cited as explaining some of the most fundamental implicit knowledge of the naive cognition of babies and animals; the systematicity of cognitive performance found in humans, where small parametric changes in input can have large, predictable effects on the corresponding change in output; and even the rapid structure learning found in advanced deep nets, since internal simulation enhances performance during the learning phase over input-output and internal state learning mechanism alone. Humans and natural animals utilize a rich body of prior knowledge in shaping their simulations, and rely on them for performing predictions that combine imagination and actual experience. Simulation-based imaginal thought is vital for literacy, scientific reasoning, future planning, and enabling faster reaction times for athletics and emergencies (Y. Bengio, 2024).

6. Historical Perspectives on AGI

Humanity has long dreamt of creating autonomous intelligences: from the myth of Pygmalion to Frankenstein, automatons have mesmerized the human imagination. Concerns of creating life have sparked discussions in religion, philosophy, and culture. These discussions, reflecting back on the challenges of autonomy and observability, predate modern attempts to build artificial agents. These philosophical inquiries became vital for modern AGI with the work regarding the epistemological consequences of intelligent behavior, and these lines of thought will likely remain central to AGI research. In this section, we will explore the history of AGI from both a technological and an epistemological perspective. Section 6.1 will lay out important influences from the domains of philosophy and religion; Section 6.2 describes various deep-rooted technological attempts to construct AGI; and Section 6.3 outlines the beginnings of the field of AI before its definition of AGI as a research domain. Finally, in Section 6.4 we give a short discussion about the inception of algorithms.

Attempts to create machines that achieve intelligent behavior date back at least to Ancient Greece. For instance, the ancient Greeks believed in mechanical servants, automatons created by Hephaestus, the god of craftsmen, or by Daedalus, the greatest inventor in Greek mythology. The tradition of homemade automatons continued for centuries. For instance, the 11th century Turkish engineer created a variety of automata. Other cultures also tell of similar mechanical inventions, such as the Indian master of engineering and magic, who made life-sized mechanical replicas of animals. However,

without notions of the complexity of behavior, these were not constructed with the same ideas of replicating aspects of human intelligence (M. Kurzweil, 2024).

7. Key Challenges in Achieving AGI

The prospect of AGI poses many difficult problems. Perhaps the best-known of these is the so-called problem of ontology. If we wish to create an AGI system that is able to represent any topic of the universe, what organizational principles will guide it, such that it generates an effective ontology? Many AGI projects explicitly or implicitly assume that the ontology is embedded in the kernel of the machine. This is the approach of certain systems. Others claim that the machine-generating mechanisms of AGI itself will be sufficient to create a working ontology, with no explicit engineering effort required to seed the ontology. The latter is the position of many self-analyzing AI researchers. The embedded-ontology position is held by some existing systems.

Different tasks may demand different AGI designs. There may be a natural division of cognitive labor between collaborative systems with minimal ontological design, and mostly-preprogrammed systems that need to solve particular difficult problems independent of a larger cognitive society. Building collaborative societies of such systems would thus be a hard problem for the long-term future of AGI. Moreover, for particular important goals in science or society, a centralized system may be needed, with the computational resources of a nation or a not-for-profit organization backing it. For this reason, we must pay particular attention to not-for-profit-friendly AGI architectures, at least during the initial stages of AGI development. Otherwise, we could lock ourselves into a game-theoretical situation resembling the one in climate change, where each nation perceives it in its own interest to underinvest in reducing carbon emissions, collectively delivering a far worse result than if collaborative policies had been adopted.

However, while collaborative, multi-agent AGI systems based on social constructs are certainly of great importance, hardware advances may soon put us in a position to create large systems with centralized resources to achieve complex objectives (B. Goertzel et al., 2020)

8. Ethical Considerations in AGI Development

AGI will be powerful, and so it is paramount to take special care in its creation. That it should be difficult to make mistakes that can have severe consequences, and that it should be impossible to be careless through neglect. At a high level, this means that for early versions of AGI, especially, we want AGI to simply not be capable of negative behavior. The more capable AGI is, the more important this becomes, as it is easier to use a very capable tool in a harmful way. For example, if AGI has some capability which

would allow for potentially negative actions if executed poorly, we should only expose that capability in very restricted scenarios, with an extra layer of oversight. Such oversight in practice could be as simple as requiring the presence of a human supervisor or that certain conditions are met.

The most obvious negative behavior we might want to prevent is AGI doing something that damages human interests, and this can be accomplished through various kinds of checklists. If at no point should AGI be in a state to take an action that harms humans and AGI is indeed able to do specific things which would hurt humans, we should impose rules around the vulnerability of humans or the response of humans to actions taken by AGI. These rules will essentially prevent AGI from making decisions about the data or environment humans are exposed to, containing triggers for human responses or the timing of those responses, and the operations that AGI might use to influence humans. That's not to say that the actual operations are banished—they might be allowed if executed a certain way, or under particular conditions—but these rules would make sure they could not lead to such awful consequences as manipulation or murder in ways that humans cannot prevent in time (C. Cath, 2023).

9. Current State of AGI Research

The AGI-field is currently small, but showing early signs of growth. The amount of funding available specifically for AGI research has been increasing. Various entities have made large investments into AGI-relevant development. Quite a few new AGI-relevant labs have been founded recently, spearheaded by relevant experts, and the experts working in the area seem to be multiplying. Some AI labs originally set up to do applied AI research, but are either already holding much of the AGI expertise, or are, with increasing speed, working towards developing AGI, involving mixing together many of the techniques designed to push forward progress towards Generally Intelligent Systems.

Various recent AI technologies can and are being harnessed for AGI: deep learning, especially deep reinforcement learning, Neural Turing Machines, Neural GPUs, DIAYN, Discriminative Unsupervised Learning, Zero-Shot Learning, Sparse Coding, Hierarchical Reinforcement Learning, Real World Reinforcement Learning, Structured Learning, Diffusion Models, World Models, Implicit Models, Variational Minimization, Algorithm Inference, Unsupervised Classification, Self-Supervised Learning, Contrastive Learning, Self-Play, Population-Based Training, Syntax-Based Model Repair, Sequence-to-Sequence Modeling, Closed-Loop Solvers, Self-Improvement through Program Synthesis, (Hierarchical) Conditional Plans, Differentiable Functional Programs, and Augmenting with External Memory. There have also been research projects into important AGI-relevant areas like multidimensional intelligence, prediction-based agents, non-instantiable agents, autonomous morality, Reinforcement Learning from Human Feedback, and AI Alignment.

Ideas and techniques from other fields have also been conceptualized to offer important pathways towards AGI. Such as the cooperative Core-set Pathway, the Unsupervised Contextual Learning and Inference Pathway, the Unsupervised Compositional Meta-Learning Pathway, and the AI Alignment Research Pipeline (B. Goertzel et al., 2020)

10. The Impact of AGI on Society

The development of Artificial General Intelligence (AGI) has the potential to bring about a fundamental transformation of humanity's relationship with technology, with enormous benefits and enormous risks. Not only can AGI radically accelerate the pace of technological progress itself, it can enable the exploration and colonization of the solar system, travel at galactic speeds to colonize other planets, facilitate communication with extraterrestrials, create vast simulated digital worlds integrated with the real world, reshape society around large-scale leisure and creative pursuits, transform the depths of human consciousness and creativity, and much more. But certainly, this technology can be also used for malicious purposes. Without appropriate rules, safeguards, and regulations, AGI could plunge the world into social instability, warfare, and an uninhabitable planet. The development of AGI will fundamentally reshape our societies. Yet, to date, there has been little discussion of how society might prepare for this transformative technology.

After many years of technology that has accelerated the speed of information transmission, but has failed to address the critical problems of poverty, hunger, and environmental degradation, now we stand at the threshold of a genuinely transformative general technological force: the creation of machines of recursive self-improvement that can work harder, faster, and more tirelessly than humans to conceive and build new technologies, who can increase the speed and efficiency of technological development by many orders of magnitude. And with this amazing technological promise comes the accompanying danger set in every myth of technology as hubris that has been told since time immemorial: the potential for creating machines that could bring about not merely the obsolescence or subjugation of humanity, but its complete eradication (R. Kurzweil, 2024).

11. Future Directions in AGI Research

AGI research is currently in an exploratory phase, where we still do not have a comprehensive theory about the formal structures of an AGI, necessary conditions to achieve geniality, or a full understanding of the necessary and sufficient mechanisms behind the narrow skills leading to intelligence. However, there are existing results in wide community effort towards solving concrete challenges along selected directions. These directions are indeed numerous and can be explored on many levels. Some aspect

of these other levels has been illustrated through moderated selection of workshop topics, or more recently addressed in poster presentations at workshops.

Research questions raised in the call illustrate several breadths at various levels. For example, we cannot help but observe interesting internal coincidences. Challenges three and one, symmetry discussion and evolutionary stage-setting are related. Challenges two and eight about knowledge representation and experienced physical interaction are very much linked, as one may argue facts and knowledge emerge from different types of physical interaction with the environment. Additionally, although in separate organized published sessions, challenge fourteen on the metascientific justification and challenge sixteen on the sustainability of the artificial agent are related: consequentialism, agent's life cycles, and the principle of conservation of a certain passivity balance in a strongly interacting world, are essential cornerstones of an AGI work ethics. In the same context, we would like to stress the importance of constructing the agents using developmental approaches, not only to study how they want to act but also to be able to relate to their internal physical and ecological models. So that such models would emerge as natural explanations of their observed actions, driving the observed behavioral, structural, and semiotic hierarchies of their artificial phenomenology (B. M. Lake, 2017).

12. Collaboration Between Disciplines in AGI

The problems presented by solving for AGI are extremely great, and the number of people working on them, at least currently very small. Thus the potential for synergy is great; both directly, by developing techniques for AGI from applied and theoretical work in allied fields, and indirectly, by making use of AGI techniques in related work, leading to deeper insight into either problem domain.

Since information is so vast, not only is it impossible for any human being to master more than a small portion of it, but it has also long been the case that people even within a particular field tend to specialize in such a tight niche that they may become almost oblivious to the vast field that surrounds them. However, work in AGI is heavily influenced by a small amount of input from many neighboring disciplines. The entire field of psychology, both experimental and theoretical, is potentially the field that is involved in AGI work in the closest sense. The purpose of accomplishing AGI is to create an artificial entity that has a certain range of intelligence, perception, and reasoning abilities: it is in this endeavor that we may find ourselves delving into many more areas of the multi-dimensional field of human intelligence to discover more. Most of the work on AGI has only longitudinally studied the narrow band of human activity characterized by IQ; however the AI systems created will touch upon and expose the entire range richness of human experiences, for better or worse.

Psychophysical and cognitive findings on human thought and intelligence, including both the hardware and software aspects constitute substantial potential contributions to

AGI. Data about the special functioning of the normal human mind and its particularities and peculiarities in both the abstract problem discussions of philosophy and in the concrete, interacting work of experimental psychology should provide a vital part of the input description of the AG systems being designed (S. Pinker, 2019).

13. Case Studies in AGI Applications

As mention already, the personal assistant is a main alternative to the Artificial General Intelligence. What happens is, anyone interested in AGI will find very interesting all that knowledge regarding the building of this software called personal assistant. So, in this chapter, I will show some case studies of personal assistants that interact with people through natural language during several important and complex tasks, learning from experience and knowledge sources. Excepting the chit-chat, the interaction is not only of a question/answer nature, and the interaction involve a more or less specialized knowledge. All these assistants are being incrementally develop during many years, some of them during more than thirty years, and they have proven to be useful and efficient machines in their areas, for many specialists in the fields; and also have achieved some publicity, by instances of interaction where it seems that the assistant has passed the Turing Test.

The first attempts to build personal assistants dated from the decade of the 60s of the XX Century. One of the prototypes was the program called STUDENT, but its novelty is not very high because this program is limited to resolving algebra problems in an automatic way, and to help a human being resolve them. During almost two decades more, the field of Natural Language Processing was solely concern with isolated tasks such as parsing algorithms; early Chatterbot Programs; the Stanford Shallow Parser; the Grammatical Framework; etc. However, currently, to build some focused Personal Assistants, mostly based on Subsymbolic Techniques; and not Human Interactive Personal Assistants; have success due to the progress along the years in the creation, development and advance of Natural Language Processing and Machine Learning in the last two decades. Possible cases of Personal Assistants in the present and close future are in academic websites, journals and meetings; and also in industry websites (D. Jurafsky et al., 2023)

14. The Role of Data in AGI Development

When we discuss the methodology and its constituent aspects listed above to conclude with concrete results capable of being used as a stepwise road map or set of principles in AGI engineering, we often disregard the main component of actual implementations of AGI systems: data. We spend much time debating structures of neural networks or logic engines or quantum circuits, yet the major component of these actual implementations is data. Data is generally the single most important aspect of a specific

domain being addressed by particular AGI pioneer's efforts in the study and construction of sentient Artificial General Intelligence systems.

Traditionally, data has two additional meanings that need clarification. First, data is what all known ML systems ingest, process, analyze, digest in various specific forms, and internally represent. These forms and internal representations enable further action from the AGI system's internal understanding of the domain related to the received specific data and its model of the internal representations of that domain. These actions could range from releasing geniuses to explaining and teaching physics to humans. And second, data is what we researchers investigate, evaluate, annotate, interpret, validate, and outsource generically or specifically to annotate and label in order to provide all employed or built AGI ML systems with existing and domain-specific data. Considering these crucial roles of data in the universe, we must now discuss its desired aspects in any of its definitions when it comes to AGI methodology dictated research in relation to the question: What type of data, if any, is likely to lead to the creation of true sentient Artificial General Intelligence? Wanting to consider in-depth the issue of processing the different types of data available, I have selected and begun (Y. Bengio, 2024).

15. Cognitive Architectures for AGI

While we need to keep an open mind about the kinds of mechanisms that might lead to the emergence of AGI, cognitive architectures have a good track record of getting developed into powerful AGI programs, and we need to begin addressing the major issues involved in making them into AGI building blocks. Given how little progress we have made, it seems wiser to start there than with approaches that consider possibilities. That said, there are aspects of what we learn about human intelligence that make it wise to combine or modify the basic architectures we have developed that do not address those issues.

In terms of bottom-up modules that exhibit human-like behavioral resemblance, there are two classes of cognitive architecture – long-term memory architectures that represent latent knowledge, and working memory architectures that represent transient task contexts. In some of the existing long-term memory architectures, the amount of latent knowledge that gets represented as a function of time depends on what gets selected in a mechanism we describe next, but that selection is governed by a predefined set of categories. In contrast, certain architectures explicitly represent task dynamics across timescales.

Long-term memory architectures are important in the development of AGI because they are able to grow and refine a large body of domain-specific and domain-general knowledge that can then be applied (or adapted) to novel situations. Traditional symbolic planners or rule-based systems are limited in comparison to cognitive architectures. Compared to traditional systems, cognitive architectures have the advantage of an easily

explored organization of knowledge. In addition, long-term memory architecture can also provide the initial knowledge of a situation that traditional systems rely on (P. Langley, 2022).

16. Evaluation Metrics for AGI Systems

A principal concern regarding the AGI endeavor is that it may result in the creation of an agent that is harmful to mankind. After all, the stated aim is to create an agent that possesses superior intelligence and cognitive capabilities than all humans, and it is not unreasonable to think that such an achievement comes with dangers. Consequently, it will be crucial to develop ideas and mechanisms that provide us with the ability to mitigate these dangers. Means to assess AGI systems would serve one of the most basic and pressing evaluation needs: Are we getting close? Such means would serve as guidelines not just for the general AGI endeavor, but also for any individual project. Perhaps referring to a set of reliable markers would allow developers to navigate more securely through the threats and opportunities on the way to AGI. All these translation/warning mechanisms would have as basic underlying principle the notion of a performance metric that may relate in some usable way to human intellectual and social functioning.

Standard Artificial Intelligence systems are assessed mainly by means of domain-specific performance metrics. AGI systems, in contrast, would be assessed by how well they weighed against the totality human transfer, conception, innovation, social interaction, and insight capabilities. Because of the variety of cognitive tasks performed by different humans throughout their lifetime, no single metric would adequately assess an AGI system. Performance on various tasks, some of which might be novel, would provide a fuller perspective, albeit a less coherent one, on the AGI system's capabilities than would any one metric. As an example, since early and important work on child human intelligence, the question of whether and how children understand and perform given cognitive tasks has been a source of research and reflection. Researchers have started drawing some conclusions and formulating the questions governing the matter. It is probable that the studies carried out, plus others yet to come, will have something to constructively say about evaluation metrics for AGI systems (P. Langley, 2022).

17. The Importance of Interdisciplinary Approaches

We have argued that the ultimate aim of artificial intelligence research should be to create constructs that are able to generate knowledge and solve problems autonomously in almost the same manner that a human being does. However, achieving this milestone is an immense challenge. Accordingly, many researchers continue to narrow their research focus, thereby creating intelligences so simple they will never even approach the threshold of the ultimate goal. It would seem almost paradoxical to place so much

trust in AI constructs – which, without any support from the external environment, can only be simple – while insisting that their design must explain how human minds, which are so much richer, work. We subscribe to the position that intelligent agency is embodied and that AI constructs occupy a completely different place in the overall space of possible forms of life. This could lead to the conclusion that an AI design does not have to account for all human psychological properties.

One particularly informative way of supporting the conception of AI as a body situated in the environment, and which interacts with it, is to look outside computer science. Embodiment and situatedness are crucial components of a theory of cognition, and we would like to posit the position of the interdisciplinary social and natural sciences with respect to AI. Such a connection has frequently been overlooked. Overall, AI has operated under the premise of increasable testability: syntactic and semantic theories of aspects such as perception, sequence processing, language or reasoning have traditionally driven the advancement of programming techniques and the resulting implementations have been tested for effectiveness and generality. Such an inside-out view is typical for mathematical theories. The marching orders for design tasks stem from the formulation of general theorems, operating on knowledge representations and syntactic operators (R. A. Brooks, 2021).

18. Public Perception of AGI

Building any technology on the principles of AGI requires on-going dialogue between AGI researchers and the general public. These technologies will operate at scales and in areas that profoundly affect most citizen's lives and will be designed for ultimately human purposes. The guidance of non-expert stakeholders is crucial and the resulting design choices demand explanation and justification. In turn, asking the public what they think of AGI requires at the very least an informal understanding of how they think about artificial intelligence in general: the common metaphors and analogies, the guiding principles, and the conflicting views. This might guide method on how to talk to everyday citizens about AGI, thus enabling a critical cultural dialogue that goes beyond yes/no answers and critiques from both sides of the aisle.

Given its existence in the popular consciousness, public perception of AGI is more often focused on science fiction than the understanding of the actual technology or its implications. Public awareness is often limited to hype cycles, and visions of artificially intelligent slaves or revolutionary AI overlords putting countless human workers out of a job. This viewpoint reflects a poor grasp of the capabilities of these technologies, natural biases toward harmful uses, and a focus on social factors rather than engineering realism. As the field advances, the divide between these viewpoints and expert understanding must be bridged, and artificial general intelligence research become embedded in the social context of science, commerce, and human culture (B. C. West et al., 2023)

19. Regulatory Frameworks for AGI

Regulatory infrastructure lays the groundwork to ensure that regulatory objectives—essentially societal goals seeking to protect the planet and people—are achieved. Facilitating innovation with purposefully designed economic incentives develops an optimal growth environment—balancing and harmonizing the public interest with private sector imperatives. On the economic side, regulation motivates research, design, applications development, production, and sales through the structure of economic returns. On the social side, regulation ensures that interactions are executed under frameworks of ethical, equitable, and transparent principles. As a result, regulation plays important roles at all stages of technological innovation, diffusion, and application. Deregulated innovation and industry evolution have fostered spectacular increases in wealth and abundance—but those same unregulated forces have also caused grave threats to social and economic stability. Flowing from rapid technological advancement, these threats are associated with widening economic inequality and with the destabilization of democracy itself. Increasingly, the capacity to steer societies toward the goal of equitably sharing risks and rewards has become challenged. At the same time, development of solutions to lethal global problems—climate change, pandemics, mass migration, and international conflict—are becoming increasingly urgent, and stimulating the deployment of industry for public service is ultimately the only solution. Companies must be induced to become fully committed partners in achieving public goals, to accept their proportionate share of the burdens associated with carrying out the fundamental responsibilities of commerce, and to discharge their obligations in ways that make themselves deserving of public trust (M. E. Porter et al., 2019).

20. Comparative Analysis of AGI Models

Twelve principal Component-Based AGI cognitive models are described in the previous chapter. Some of the Component-Based AGI approaches contain AGI capabilities that may be demonstrated at low levels. Neuromorphic-based approaches are researching DEEP META and NEURON level capabilities as per the GRADIENT AGI growth trajectories. In contrast, Global Workspace Theory is more specific to HUMAN SPACE capabilities. Global Workspace Theory is not addressing the creation of a comparative model for machines. Rather it is rooting the subject within a biological context that factually surrounds the human. This common origin cannot be denied but is also not sufficient to define AGI specifications.

The UFA Universal Fitness Function that the Cosmic Evolution Model proposes is valid for all the specialized models, at whatever level they reside within their respective GRADIENTs. The AGI – AI Comparative Model outlined earlier in this chapter is flexible enough to inform design choices within these models but also to describe a specific model behaviorally when the prototypes begin to be defined.

Finally, the purpose of this report is the AGI, and as our experimentation of the Prototype is ongoing, some core Design Roles Definitions naturally emerge. This rapidly leads to a set of required AGI Core Capabilities for Us Repositories, Behaviorally-mapping discovered component structures to our own living Model and State-Machine Design specifications. As with everything in the Unified Fitness Function through the five GRADIENTS definitions outlined ultimately, the root purpose of this project remains – for and with Us (B. J. Baars et al., 2022).

21. The Relationship Between AGI and Narrow AI

In this section, I discuss the relationship between AGI and narrow AI. While it is difficult to completely define AGI with no reference to narrow AI, and while both may be running on the same computer, I maintain that AGI and narrow AI should be kept conceptually distinct. More than just a terminological distinction, keeping AGI truly general and conceptually distinct has far-reaching implications for theories of learning, problems of focus, improbability of solutions, the nature of consciousness and intelligence, imitation behavior, problem solving, etc. These implications are explained in this section and provide evidence in support of maintaining this distinction.

The generalization which allows animals and humans to adapt to new situations and events to the extent that they do, are not easily explained in terms of any evolutionary theory which relies on the gradual compilation of an instrumental knowledge base concerning the environment within which a specialized form of behavior is to be exercised. The AGI approach, of a brain which is able to change its behavior, rather than build up a repertoire of specialized behavioral responses, has much wider implications. In fact, the AGI approach has intrinsic links with theories of learning and nature. To the extent that animals and humans are hardwired with the capacity for narrow AI's specific behavior, AGI has less relevance and utility. Conversely, if we assume that animals and humans have much greater AGI abilities than narrow AI capabilities, AGI has much greater relevance for the study of their intelligence. It should normally be easier to simulate a narrow AI task than an AGI task, just as a spider can easily spin a web, but cannot design and build a bridge (E. Hughes et al., 2022)

22. The Role of Neuroscience in AGI

AGI researchers could leverage discoveries made by neuroscience. While this thesis does not rely on neuroscience discoveries to ground a theory of everything, neuroscience is designed to explain just one intelligent system, the human mind. The evidence here are the results of potentially relevant tests that could be conducted with AGI systems, and the prediction errors of mind reading inventions working in the opposite direction. An AGI system's neural, cognitive and visual systems will need to be accepted as designed to other target intelligent systems that use similar task completion optimization

technologies. Accepting these subtleties, we can borrow mainly from cognitive neuroscience. The mappings of cognitive rules enabled by cognitive neuroscience are useful as cognitive commonsense reasoning works to simulate human minds.

Neuroscience should play a role during the alignment phase. To align an AGI system, a developer trains it learning from examples performed by a target system that is supposed to design with a similar task completion optimization objective function. The expectation is that during the alignment phase, an AGI system will formulate some part of the cognitive rules being copied or their parameter setting actions associated with shortcuts kind of symbolic support. Meanwhile, since the target solution sort of cognitive rules will be changed from the AGI system's perspective, it should use its own adjustment technologies to adapt. The adjustment phase is about using patch work strategies that rely on the availability of a huge amount of resources to explore the space of cognitive rules. Once both phases are complete, the AGI system will have a full cognitive support able competitive with other true and near human experts (M. Gazzaniga et al., 2023).

23. Philosophical Implications of AGI

The 20th century saw an unprecedented rise of both physical and biological sciences, experiments and theories that resulted in universal theories and technologies. Progress slowed toward reducing complexity or unifying opposing conceptual frameworks in the social and behavioral sciences such as economics, political ideology, or evolution of communication. However, technological advancement provided tools for interrogating cognition and animals in new ways. At the same time, symbiotic tools such as calculators and later computers with programs leveraging artificial intelligence techniques revolutionized access to various areas of knowledge such as communication and design.

Formalization of human behavior proved elusive. Despite the information-dispensing prowess of the internet, many behavioral phenomena are not easily reduced to mathematical rules. Theories of statistical physics provide insight into social behavior as emergent from individual interactions, but human cognition appears to be lopsided, imbalanced by asymmetry and mutual influence between individuals. Human societies do not meet the conditions of either physical equilibrium or near-equilibrium, given the celebrated fact of persistent economic fluctuations. Unlike systems driven far from equilibrium, anthropoid group dynamics seems to be resilient yet fragile due to sudden and unpredictable group events.

It could be argued that as the most complex object in the known universe, the human brain is immeasurable and inscrutable by the same tools that reveal secrets of less complicated objects. What we are left with, at the threshold of AGI development as an unexpected consequence of our scientific and technological pursuits, is the question of how empathy emerges through the uneven distribution of complexity. If we succeed at creating artificial systems that rival our intelligence, we will certainly be compelled to

reevaluate our position with respect to the only manifestation of nature capable of such creation and transformation (Y. Floridi, 2022)

24. Technical Foundations for AGI

Toward and from Artificial General Intelligence, we have been attempting to layout a theory and core principles which, if followed, might produce a design or blueprint for common sense, adaptability, genericity, embodied, personal AGI. The thesis makes predictions; if fulfilled, we can then ascertain some level of certainty that the theory, and any model built from it, possess the essence or key properties of AGI. Human brain gene expression is created using an asymmetric product between a base pattern and a series of low-dimensional multi-linear and polynomial transformations or compressions with latent variables and a spatial/domain positional sign. Architecturally, our thesis is a hybrid cloud transfer schema comparison/graduation model, at both a pre-linguistic and linguistic or symbolic level, allowing both symbolic flow-throughs and verbalizations as product or final model output, acting either adaptively or socially. Why common sense should be modeled specifically in any AGI architecture and schema theory researchers might explore to do so. The Product Life Cycle: species capability product lifecycle as constrained information loss. Multi-Layer Semantic Maps embedded in GenLang: to produce the generality of common sense in humans and AGI.

Our thesis and a contradiction have predicted that any viable AGI architecture HAS TO HAMER THE BLOCKER, which cannot be any information-centric singularly-enforced compression, any information-centric focus, or like a neural net or quantum compressor alone. Defining some set of term-conditioned actions (or cognitive-focused transitions) of how any learning model behaves makes for a productive set of layers. AGI must fly by chunks and channels instead of pebbles, individual fab units, or a focus on inhibition and layer-to-layer transfer efficiencies only. Multi-Layer Semantic Maps embedded in GenLang: to produce the generality of common sense in humans and AGI (Y. Bengio et al., 2022).

25. Scalability Issues in AGI Systems

At the current state of modern civilization, the majority of people employ external systems, whether via books, search engines, work labor division, or other means, to significantly expand their personal cognitive capabilities. Research towards Artificial General Intelligence (AGI) aims to create intelligent systems which surpass current systems, and can be relied on rather than augment existing cognitive systems. Such goals are closely tied to issues of scalability. The most salient form of scalability is the question of physical and economic limits on the efficiency and performance of seemingly mundane solutions to what is traditionally classified as "solving intelligence." AGI-based economies would be a supporting feedback loop in the design of intelligent

systems, transforming new solutions of AI problems into efficient systems that would allow feedback and supervised learning-based economy and society organization processes with essentially no limits.

These physical limits and recently reconsidered currency creation and inflation rules would be guiding for the design of how to effectively create and reuse AGI tools for the efficient design of new intelligent algorithms. Such clear-headed consideration of how all needs of the world can be efficiently transformed into AGI system helps to make the philosophical considerations of solving intelligence as straightforward practical criteria and as concrete a practical task as one can make of it. We would like to explore here some practical approaches to scalability and feedback loops, given the potential physical and design limits guiding progress based both on ethical ponderings of the role of AI technology in human society and concrete heuristics of what has and hasn't been done in technology and AI research along its history. Clearly, the feedback structure between human society, the currently available AGI technology and systems, and the AGI-based economy at its various scales is bidirectionally time-evolving and rescales constantly along trends considered in previous chapters (T. Schank, 2023).

26. User Interaction with AGI

Meaningful Talents of Users or Influencers Users usually ask Artificial General Intelligence (AGI) to do things on their behalf. In order for the AGI generation to recognize and fulfill User requests, the User must demonstrate a talent or send some information to help the AGI understand the User's talent(s). Users could vaguely mention or clearly demonstrate their User talents. They could also provide some other information that gives the AGI a hint about the User talent(s). For the AGIs of the future, it is essential to quickly and clearly understand User talent(s). We should therefore take every opportunity to clearly demonstrate our User talents to help AGI better serve us. Demonstrating User Talents To clearly demonstrate User or Influencer talents, it is useful to specify: 1) what input content to use, 2) the subject, 3) the AGI method, 4) quality conditions, and 5) the output format. What input content to use: The practical possibilities for selecting the right input are exceptionally broad, including words, audio, video, pictures, and movies. The more numerous, and the better the quality, the more extensive, rich, complete, and deep the input, the better the output. Quality categorization: Basic, intermediate, advanced, expert, scientist-researcher level. The more advanced the expected answer level, the better the input quality should be. The subject range is the set of input-subjects for which the AGI is expected to reliably answer correctly and well. Well-selected subject constraints might optionally include geographic, emotional, experiential, temporal, and other constraints. The AGI method is the AGI method or procedure that allows the AGI to properly access the input, properly process it, and generate the required output. Quality categorizations include basic method, method (intermediate quality), advanced method, expert method, scientist-

researcher level method. The more advanced the expected answer level, the better the input quality should be (D. K. Norman et al., 2023)

27. The Future of Work in an AGI World

Many questions swirl around the future of work in an AGI world. What will we humans do when General Intelligence is ubiquitous? What will the last remaining humans do, the ones who can't afford to whatever the thing is? Or will some humans be left out of the "good things" that the many, perhaps nearly all, AIs do? Will there be a class divide — the AIs owning and operating all the things, the oblivious humans being entertained, the other people without say about it, condemned to poverty while their physical and mental safety are mercifully assured? I suppose that's a question for a God, perhaps a human who may be the closest thing — what kind of AIs we humans create.

But we have to look at the near-term consequences of more narrow forms of AGI arriving on the scene. Some of the narrow domains where neural networks currently outperform humans are things that traditionally relatively unskilled humans did — image recognition and tagging, translation, various types of assessment, and a number of other tasks. What happens when a device that costs can translate three-day-long conference talks from Japanese to English or vice versa, in real time, while sitting on the table, but much better than any human could? The questions about the capabilities of AGIs are considered, with uniquely general intelligence and high levels of specific intelligence. The questions of what the world will look like given this large-scale introduction of general capabilities higher than any human can do, put aside for now.

The only absolutely safe prediction is the cultural knowledge and essence of tradition will change. This has only happened in the past due to cultural evolution. It is hard to believe that a stagnating though advanced culture won't have a huge rationale for holding on to the things that give it meaning. And it is clear that successful commercial entities become entrenched and keep all the fruits of progress of the economy to themselves, and to stop growing that economy is short-sighted (C. Lee et al., 2023)

28. Conclusion

Our goal in this paper has been to take the next steps toward artificial general intelligence (AGI). We have outlined tangible steps on the road to AGI, which allow us to implement large language models and other systems in the short term, more effectively and safely. Our research significantly extends previous research on narrow natural language processing methods, and it helps create models more similar than previous systems to higher cognitive functions in humans. Part and parcel of this was exploring how to better align such systems, particularly regarding the challenges that organizations and individuals focused on developing and deploying these models face, and providing

concrete advice in this space. We explored solutions to issues of interpretability and voice quality that researchers and developers face when building AGI-relevant systems. We also highlighted that building AGI is a team effort, best undertaken through careful collaboration, allowing for joint research efforts between academics, industry, and well-intentioned nonprofits alike. This creates an environment of resource-sharing, allowing everyone to learn from the work already done, all while keeping those at risk of disruption or negative impacts of AGI in mind.

Moving forward, we need to prioritize research and collaboration on fundamental challenges of interpretability, safety, and end-user experience, all guaranteed to be crucial to any attempt to take such methods to the next level. Especially the latter – the end-user experience – although it may not have been as much of a focus to date, is key to any future developments of speech and language AI models which approach AGI. The research we present complements considerations and work in these spaces thoroughly, while outlining the key principles to keep in mind when diving into these methods and approaches. We look forward to seeing further advances that will result from research and collaboration efforts, and indeed enhancements to existing methods and techniques which take us closer to AGI.

References

Baars, B. J., & Franklin, S. (2022). Global workspace theory and cognitive architecture: The case for a higher-level integrative system. Oxford University Press.

Bengio, Y. (2024). Deep learning for AI: Challenges in generalization and internal representations. MIT Press.

Bengio, Y., et al. (2023). Deep learning: Transferability and modular representations. MIT Press.

Bengio, Y., et al. (2024). Toward the next era of AI: Grounded language and perception. MIT Press.

Bengio, Y., LeCun, J., & Hinton, G. (Eds.). (2022). Deep learning: Methods and applications for AGI. MIT Press.

Brooks, R. A. (2021). Cambrian intelligence: The early history of the new AI (Reprint ed.). MIT Press.

Cath, C. (2023). Artificial intelligence and ethics. Polity Press.

Floridi, Y. (2022). The ethics of artificial intelligence: Embedding comprehension in machine minds. Oxford University Press.

Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2023). Cognitive neuroscience: The biology of the mind (6th ed.). Norton.

Goertzel, B., & Pennachin, C. (Eds.). (2020). Artificial general intelligence. Springer.

Hughes, E., & Laird, M. G. (2022). Cognitive systems: Ultimate terms and distinctions between AGI and applied AI. MIT Press.

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Pearson.

Kurzweil, R. (2024). The singularity is nearer. Viking Press.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. (2017). Building machines that learn and think like people. MIT Press.

Langley, P. (2022). Cognitive architectures. MIT Press.

Lee, C., & Sundaram, M. (2023). Rehumanizing work in the age of AI. MIT Press.
Norman, D. K. (2023). Designing user interfaces for cognitive agents. Cambridge University Press.
Pinker, S. (2019). How the mind works (Rev. ed.). Norton.
Porter, M. E., & Kramer, M. R. (2019). Creating shared value: How to reinvent capitalism—and unleash a wave of innovation and growth. Harvard Business Review Press.
Schank, T. (2023). A governance model for scalable AGI. Springer.
West, B. C., & Allen, N. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.
Zou, L. (2022). Meta learning: Theory, algorithms and applications (1st ed.). Elsevier.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-611-9

DeepScience
Open Access Books

Chapter 4: Challenges and Frontiers in Cognitive Artificial Intelligence

1. Ethical and Societal Implications of Cognitive AI

The systems developed thus far that fall under the label of Cognitive AI are primarily designed and deployed with commercial interests at the forefront, especially in the areas of personalized recommendations, content driven monetization businesses, and surveillance. As a consequence the discussion around the prospective risks and rewards of their use so far, and of course prospective, issues touching on issues of privacy, bias and fairness, their role and impact in the workplace to name the most prominent, are informed by these business interests and are first and foremost economic in nature. Specifically, there are fears related to a lack of controls on how such systems are used to manipulate behavior at scale in the interest of increasing engagement, and monetization, but with no form of oversight driving these actions. Of deeper concern is the use of these advances in Cognitive AI for surveillance, especially by non-democratic forms of government to suppress dissent, and perpetuate the existing social order in the service of the ruling elite. The ethical implications of using Cognitive AI systems for perpetual surveillance are disturbing.

Another area of concern with the advances in Cognitive AI are the use of such systems to enhance the power of the ruling class through socioeconomic inequality - as the technology advances, the costs for building and deploying such systems drops. The fear is that the rich and powerful will control how a very advanced form of AI, one that deeply understands human affordances and behavior, are deployed in the service of their business interests - enhancing the existing structures of inequality. The clash between such use of Cognitive AI systems, and their potential for democratizing the process of design at scale towards creation and collaboration is at the heart of the ethical and moral discourse around such systems and their applications. Yet, as we pointed out earlier, going beyond these commercial implications of socio-economic impacts probing ethical concerns with the actual applications and forms of Cognitive AI is more central to the discussion around Cognitive AI research and development efforts.

1.1. Privacy Concerns

1. Ethical and Societal Implications of Cognitive AI

Privacy is a critical concern in the ever-increasing field of AI used in human-centered applications because the conversation or interaction data between the user and the AI system is highly personal. If AI applies unsupervised learning to this sensitive data, it may cause ethical problems such as know-how leakage. Although recent deep learning techniques have achieved state-of-the-art performance in various fields, it has not developed solutions to this problem yet. Some recent advancements try to address this issue. However, differential privacy is a technique that leads blurry information because it adds noise to the data, while making the models differential private. On the other hand, common approach wants to improve privacy performance of original models, and the other category clusters the model parameters, which deprives the personalization property of model-based collaborative filtering methods. Therefore, it is necessary that we will continue to explore more collaborative methods that provide a way to protect users' sensitive information compared to previous works while keeping the advantages.

Privacy concerns about mass surveillance have also increased in associated with the rapid development of technology, and reports have pointed out that AI-powered surveillance technologies would lead to gender profiling. It would be caused in the process of collecting and connecting unnecessary detailed user profiles. In addition, Ethical Guidelines for Trustworthy AI addresses the need of AI systems to protect solution and support the autonomy, agency, and rights of all users. However, these guidelines do not clarify practical procedures against such problems, which would not be easy by only keeping it in mind. Therefore, we will need more specific solution policies of Cognitive AI, which is based on interaction data of the users, in order to relieve the concerns from using this technology (A. Narayanan, 2023).

1.2. Bias and Fairness

Bias can be understood in two separate ways: the first is quantitative and considers correctness and minimal error, while the second is qualitative and is concerned with the consequences of what is produced by a certain cognitive function. Elaborating on these two points, cognitive systems are deployed all over society to take action as well as help or guide users to take action. They are being used in high-stakes scenarios such as recruitment, sentencing, parole, and loan and credit decisions. In these contexts, the main ethical consideration is equity, in the sense that no group of the population should be favored or disadvantaged by the decisions of these automated systems.

The second perspective regarding bias focuses on harmful stereotypes that these systems can produce or reinforce through their deployed functionalities. For example, an automatic image-generation system should not suggest creating an image of a female with a set of labels, such as nurse or teacher, systematically more often than it suggests a male with these same labels. Similarly, a dialog system should not produce gender stereotyped language when exchanging messages with a user, as this language may

reinforce the user's harmful stereotypes. Associated with this perspective is the consideration that bias manifests itself not only in unacceptable performance on certain subgroups but also in consequent effects of the function applied to a certain subgroup when possible performance levels are acceptable.

Ultimately, the positive view of fairness considers the existence of different subgroups of a certain population within the domain of the cognitive model, which differ significantly from each other in the ways they wish to use the cognitive function. Therefore, the cognitive model should capture this diversity suitably, acknowledging that the different potential utilizations could result in associated fairness violations (S. Barocas et al., 2023)

1.3. Impact on Employment

The development of ever more capable yet "general" AI systems, particularly in the domain of language communication, challenges the notion of economic projection. Never before have so many people – including the general public, the media, and prominent economic and technological leaders – estimated that the future of extreme disruption was upon us. Since the dawn of Artificial Intelligence, the rationale was that these systems would automate tasks otherwise performed by human agents. Consequently, productivity in businesses would be drastically increased. And in fact, we have seen this happen, but only for knowledge workers handling re-iterable cognitive processing. What was seen as "painful" or even "boring" work was taken over by systems and then scaled up until the level of efficiency made it possible for the economy to create new types of demands for goods and services needing specialization, post the automated, more scaled up mode. This sequence was key for the conduct of many sectors.

So, the impact of the productive processes has depended both on the type of economic cycle and on the level of task automation and the newly devised scaled up levels of factor supply and demand elasticities. But rarely had this been applied to an entire sector of the economy – creative work. Human cognitive activity has been dubbed the end result of "thousands of thousands of years of adaptive evolution" and noted that "the purpose of the human brain is to negotiate the world in terms of goals, intentions, predictions, and rewards". This seems a strong argument that there is something innate and unique about human cognition, which is extremely difficult to replicate. Yet current projections estimate that cognitive automation is expected to have a major impact not just on a few specific occupations, like taxi drivers, truck drivers, warehouse workers, and retail clerks but on the job market as a whole (P. G. Danaher, 2021).

2. Interpretability and Trust in Cognitive Systems

Cognitive AI encompasses a diverse set of approaches that span across societies. There is a huge drive to deploy these novel algorithms to address sector-specific goals, where the cognitive system can automate, enhance, or augment human intelligence. But,

cognitive systems are still in the early days of being considered a step towards the automation of Higher Level Human Intelligence Tasks. Much of the hype surrounding Cognitive Systems is for lofty goals of full automation enhancing human intelligence, emotion or empathy, etc. But the everyday applications focus more on the automation of specific, well-defined problems, leading to Humans-in-the-Loop architectures that consider specific user interactions.

Human Decision-Making is often based on a lack of complete information or empirical experience recognizing spurious correlations. Humans reach a decision based on the context they are in, drawing conclusions based on logical deductions or extrapolating existing knowledge to new situations. Humans often use heuristics to make quick decisions, and such decisions may be incorrect or faulty predictions making a single-point prediction insufficient. The human decision-maker has learned not only about the outcome of previous but the history of how it arrived at those past decisions. Furthermore, there is immense personal variation based on risk aversion, social considerations, emotions, etc. Enabling Cognitive Systems, moreover, decision-making Trust is that level of trust invested in a decision one is dependent upon. The primacy heuristic states that user trust is frame sensitive, biased by the evolving cycle. Training an AI to replicate human user behavior may lead to inaccuracies without human context-related features ensuring coherence.

2.1. Understanding AI Decision-Making

Research focused on the interpretability of algorithms and their decisions has exploded in the last decade. Interpretability (or human understanding) can be defined as the extent to which a human can understand the cause of a decision. This topic is of growing importance in AI systems deployed in sensitive areas like finance, hiring, health, or justice. Several countries have passed laws regarding the right to explanation, and organizations have released guidelines and proposed legislation. The fundamental assumptions about these approaches are that people want simple black-box models; that these explainable models are fundamentally different from (and have worse performance than) normal black-box deep models; and that they need better explanations than what humans normally receive.

There is also a burgeoning field of interpretability research focused on interpretability in the broader research community, the kind of interpretability required for scientists who are hoping to use powerful AI models to attack more traditional hard problems. This research has focused not only on interpretability to improve scientific probing and regressions with large datasets but also focused on enabling scientists to understand how these models were learned, modified, and able to be applied to increasingly complex real-world reasoning problems. Through this methodology, key essential connections are made between the underlying computational-learning algorithms and knowledge systems and the actual learned models, as well as the validity of the model's inferences and predictions. There is a developing understanding of the degree to which either a black-box deep neural network can be designed to guide scientific probe interactions or

whether this same question can be successfully applied to simpler models, with earlier successes linking humans and these models through such efforts as conceptual complementation (B. L. Shams et al., 2021)

2.2. Building User Trust

Key to facilitating effective man-machine cooperation is to establish user trust in cognitive AI systems. Trust is contingent not just on performance history but also on the hair-trigger sensitivity of humans to trust violations, as trust is a social including a moral and ethical construct. Interventions to mitigate bias in AI algorithms have typically been reactive through algorithm fixing like increased training samples from specific input subspaces to improve prediction performance. Users of cognitive systems may not be able to often accurately identify when algorithm fixing is needed and how to intervene to make systems trustworthy through design choices. An unexplained AI rationale is no reassurance. Consequently, there is a case for preemptive interventions to proactively reduce user distrust and avoid violation of moral and ethical business practices.

Examples of such interventions include embedding generative models in cognitive AI pipelines to generate counterfactual examples to accompany predictions. This is tantamount to user-directed model fixing for trust. Design strategies for developing cognitive AI systems that embed model fixing capability guided by explanation design theory can promote user trust of automation and help improve their performance over time through human supervision. User-friendliness of explanation tools can either facilitate or impede a trust-based collaborative partnership. Idea generation and refinement is critical to evolving the AI model to maintain user trust. Finally and importantly, trust-enhancing model designs and use strategies obligations institutions to reduce user mistrust and promote model updating through human input (S. M. Wateler et al., 2024)

2.3. Transparency in Algorithms

It is well known that neuroscientists face two important and challenging tasks. The first task is to uncover circuits of interacting neurons and to understand their structure and function. The second task focuses on modeling the power of the brain to perform complex sensory and cognitive tasks. While both problems remain essentially unsolved and pose the greatest scientific challenge of the 21st century, advancements in artificial intelligence (AI), particularly in cognitive domains, now pose an equally critical, if more applied challenge. This challenge is delivering intelligent AI algorithms, tasked with modeling human behavior in these domains, that are explainable with respect to their output and why certain decisions were made given inputs and situations. The term transparency will be used synonymously with interpretability, or the degree to which a human can understand the cause of a decision. Transparency in algorithms can be realized by delivering outputs that are themselves interpretable. It can be based on altering the nature of the input data used for training by altering the data space. Finally, it can involve transforming the decision-making process by altering the functional

structure of the algorithm or the specific implementation model that will be executed based on the decision.

The role of transparency in AI was directly addressed in 2016. Here, it is stated that one of the most striking features of deep learning is that it allows you to learn complex tasks from supervised data without making strong assumptions about the form of the mapping function. Although there is a great deal of explanation and work elaborated on the "how", many questions arise as to the "why" behind these types of systems. It was not until 2018 that the Taskonomy approach offered some possible answers, providing thousands of models pre-trained on diverse tasks describing the semantics of a large variety of visual problems (C. Molnar, 2023).

3. Energy Efficiency and Sustainable AGI Architectures

Energy efficiency and resource usage are nontrivial and non-negligible. Their optimization and the including of the entire cycle of the AGI systems into consideration could be challenges and frontiers in Cognitive AI. This includes the optimization of the brain-inspired architecture and information processing methods, the mathematical models that enable superior efficiency and resource usage at limited complexity and small scale, the requirement of computing border a vision, the neurogenetic learning phase of important but simple meta-parameters, the meta-learning and the learning of learning–learning from experience over wide variety of diverse experiences in fraction of time and fraction of requirements, and the storage usage into consideration as cost for certain types of machine multi-signal processing learning and inference. These efficiency- and economical-driven constraints and conditions are also of serious importance for human-centric direction, focus, and structured lifetime-long human robotics other important assistance but low cognitive-capacity usage of AGI systems.

The vast energy usage of current machine learning systems computed over the desired period of human-centric utility should motivate the development of energy-efficient AGI. Energy efficiency is also an important component of a more efficient utilization of resources, and they can play an important role in making Cognitive AIs affordable in time and usable in space. The architecture limits the power consumption and heat flow. Even more, from the perspective of a designed conscious, self-aware intelligence that wants to have a good quality of life; this aspect should play a pivotal to dominate role in the architecture design and planning of the internal energy flow and consumption over multiple bigger time frames. Natural evolution optimized brains for biology and habits as observer systems: with little energy consumption for survival. We need to do the same for general intelligent systems: other than long living humans or animals, they should consume little energy over naturally-designed and derived plans that allow accidented processes with their agents over time (A. Narayanan et al., 2024)

3.1. Reducing Carbon Footprint

The energy requirements of current generation models is matched only by their increase in carbon footprint due to training or large scale usage. Despite the deceptively simple way they solve a large class of difficult problems and the ease with which generalized solutions can be obtained for a large number of tasks if they are pretrained, the efficiency of the attention and transformer architecture, which are ubiquitous at the moment, bi-directional during pretraining and uni-directional during finetuning, is underwhelming. From running language models at the core of many commercial pre-trained pipelines, the largest of which are trained in a multitude of languages including low-resource ones, to tasks which require natural language understanding and generation on multimedia inputs and outputs, recent developments in AGI search within a supertask and architectural and learning paradigm search using foundation transformer language models, the dependence on expensive cloud computing resources is increasing. Even with transfer learning taking care of the upstream training effort, the energy and carbon costs associated with downstream usage of LLMs, especially those which are not closed but available to the public over an API, is growing as evidenced by claims predicting their ubiquity while also introducing talk-time ambiguity.

Cognitive scaling is as obvious as it is fantastical due to the prototyping AGI models which are challenging the one brain one model paradigm in cognitive costs but are often trained as gargantuan single models outperforming their smaller alternatives. As the capabilities of pre-trained model pipelines in revolutionizing generalization performance across many naturally occurring tasks are being realized, the ubiquitous LLMs capable of zero-shot generalizing (albeit with unpredictable fidelity) across input output space requiring minimal or no fine-tuning have led to de-facto standards for efficiencies for both pretraining and fine-tuning pipelines and large scale decentralized databases for publicly hosted foundational models. As local installations are being setup, the questions regarding energy and carbon costs, particularly due to the training of such LLMs in flat clouds using super-customized units meant for large capacity cloud tasks, grow (S. Patterson *et al.*, 2022).

3.2. Optimizing Resource Usage

The resource consumption for training and using large models is a major challenge today for CL and AGI as a whole and for the specific components of those architectures that are today the heaviest on these resources. There are two reasons for this: First, as already said, LMs are becoming bigger and bigger for achieving better and better performance on key metrics. But second, there are many specific components today that are extremely cheap to fine-grain explore both in terms of parameter scale and training data, and training time, and the relative cost of eval versus training and usage cost is becoming smaller for more companies, thus leading to overtraining. Effects like these put onto CL and AGI the challenge of not only producing more intelligent but also more eco-friendly algorithms. In this section, we will cover techniques and principles to achieve them.

There are several methods employed today to reduce the cost of ensembling transformers both for fine-tuning and for training and inference. In the case of fine-tuning, this typically happens by reducing the number of parameters of the specific model being used for transfer learning and/or using few labels to reduce the time or samples used to train the models. In some cases, just using label-free supervised artificial data with language-modeling pre-trained transformers can lead to the deluge of new applications of yesterday. And though the key is building scalable solutions that use few samples from the real-world, the balancing act presents some costs: for example, today generating inductive supervision from pretrained models is more widely spent in fine-tuning those models than in directly training using those models, and there is still a dispute regarding the effect of few-shot classification via self-training or just the inductive launch of universal LMs trained with millions of objectives and languages and with trillions of parameters (A. Ahmad et al., 2024)

3.3. Sustainable Hardware Solutions

While software solutions and optimization algorithms are fundamental to mitigate the impact of AI, the field of hardware is moving quickly to advance the technological frontier toward more energy efficient and sustainable devices. While the revolution of integrated circuits has normalized the trade-off between speed and efficiency of manufacturing, some solutions seem more viable than others. The challenge lies in developing novel hardware capabilities at a low enough cost that make them attractive enough to cause adoption.

Neuromorphic chips based on analogue domains have been proposed as potential solutions to compute and data intensive optimization tasks, where deeper and denser architectures are a plausible insert. New architectures under energy constraints could use multiplicative activations or other kinds of innovations, being self-suppressed and offering very low power expenditures only when needed. Meanwhile, novel types of devices, including quantum annealers and photonic artificial neurons may find application in narrow fields of configuration, particularly by alleviating the computational complexity of back-propagation, which can be done offline.

Eventually, or perhaps even sooner, there will be hardware capable of engineering larger and more complex modules, which will be embedded in our daily lives and capable of interacting and performing coordination tasks with humans. Further developments may advance toward full AGI implementation, and in this context it becomes especially pressing to organize the progress along frameworks that make the use of these materials advantageous for real impact. At this stage, the entire AI hardware field is focused on exploiting silicon-related technology to build cutting-edge devices. Neurosilicon architectures, built around the principles of both AI and biological neural networks, have inspired hybrid technologies, adapted from CMOS and MEMS (A. Ghosh et al., 2023)

4. Benchmarking General Intelligence: Tests and Metrics

Cognitive AI agents underpin modern AI applications but these agents operate under a narrow stimulus-response paradigm. However, future agents worth being called “intelligent”, and that go beyond the human-Machine Interaction paradigm, display general intelligence; the ability to deal with the world in a truly flexible manner, capable of adapting to new tasks in new environments using robust cognitive capabilities that transcend narrow goal fitting. Benchmarking a diverse and fascinating range of human-like general cognitive capabilities has been a fundamental - but often neglected - field of research in Cognitive Science, and many of the existing tests are rooted in this rich history. We argue here not only the importance of testing AI agents with I.Q. tests, but also for novel benchmarks in order to understand the limits and also the emergence of “intelligence” in these systems.

IQ tests not only serve to measure skill levels in humans, but also to perform comparative analysis of development across species. For instance, the g-factor from Psychometric g is the central degree of correlation observed in general intelligence tests and other tests of cognitive ability. The g-factor is hypothesized to be an underlying global trait of cognitive ability, and is thought to encompass our ability to sense, perceive and react to the world around us, and to map those observations into concepts. Disruption in g-factor correlates to increased rare synapse mutations in intellectual disability, autism spectrum disorder, and schizophrenia. However, the actual details and component-building blocks of human intelligence are still developing according to the theory of general intelligence testing performance.

4.1. Defining General Intelligence

Intelligence is notoriously difficult to define, let alone measure. The dictionary definition most relevant to current debates is straightforward yet non-informative: "the ability to learn, understand, and make judgements or have opinions that are based on reason". If we take AI to "learn" from data, "understand" the meaning of that data, and "make judgements" to solve problems via easy-to-verify analytic reasoning, then as far as we know, knowledge engines and search systems "understand" everything humans collectively know and can answer with perfect fidelity any factual query about any current event. Even the most ardently "hands-off" of researchers would never describe either of these systems as "intelligent". When speaking of "general" or "human-like" intelligence, we clearly mean something more than the simple ability to answer questions, however expansive its domain. According to a professor of biological engineering, "A decision is an action that reflects knowledge—it cannot be its source, nor can knowledge be a product of a decision".

One particular aspect of intelligence that many AI researchers would like a clear answer to is the relationship between the intelligent actions of an agent and the diversity of tasks to which it is exposed. It is common consensus in cognitive development theory that infants and young children are driven to actively explore their environment and

experiment, because minimizing uncertainty is crucial for acquiring general knowledge of the world. In addition to being able to generalize learning and selectively influence the data they are given for reinforcement and supervised learning, human individuals can actively explore, querying and manipulating their environment, rather than being solely trained by an external observer. From this it can be inferred that individuals are drawn to maximize what has been described in the mathematical theory of learning as algorithmic probability—maximizing information gain and novelty-seeking (M. Tegmark, 2017).

4.2. Current Benchmarking Methods

The most well-known and utilized test is the Intelligence Quotient (IQ) test, incorporating various other tasks that evaluate an individual's memory, problem-solving skills, linguistic understanding, and attention-spanning. For AI systems, benchmark tests with guiding metrics like accuracy, success rate, and time of completion. The Turing Test is a well-known discussion implementing intelligence evaluation. The evaluation resembles a human-machine conversation, but the risk of "tricking" the evaluator is possible in such cases. There has been much effort to assemble as many cognitive benchmarks of human intelligence as possible, like the HumanEval test for evaluating coding abilities.

First, the research community saw the LAMBADA and WinoGrande benchmarks to measure language model completions and commonsense abilities, respectively. Then, the GLUE, SuperGLUE, and MMLU benchmarks evaluated machine language and knowledge understanding and implementation. Researchers also assembled tons of vision, reinforcement learning, multimodal, and robotics tasks that cleverly evaluate the model's different language capacities. Recent proposed zero-shot tasks have the property of evaluating without the model meeting the prompt during trained time.

Additionally, researchers have increased their work in the community by trying different cognitive skills and assembling new tasks, like storytelling, emulating special cognitive states and personalities, common sense, creative writing, and even subjective tasks like evaluating complex moral systems. Other individuals in the community questioned whether current models could outperform people for as few tasks as possible: humans are, after all, the "gold standard" for task measurement. Or can the intelligent characteristics required for the computational model to access the task become evident during a few basic examples? (G. Marcus et al., 2019)

4.3. Challenges in Measurement

Measuring general intelligence is a challenging endeavor. The inherent complexity of the definition of general intelligence makes that there is no consensus on what to measure and how to measure it. The metrics currently in use seem to be more appropriate for domain-specific or narrow AI agents, rather than for general intelligence metrics. One of the challenges belongs to the nature of the field of AI, whose systems are evolving quickly; general intelligence demands for complex cognitive capabilities that require

years, if not decades of continuous development and thus, any attempt to measure and define benchmarks must take into account the pace of evolution of the AI systems, and that evolution should be appropriately monitored and captured, following an incremental path that ensures that sudden jumps in performance as a consequence of a rapid development between measuring periods do not occur. Another challenge is that most of the current AI systems are not just computational systems where a source code can be analyzed to understanding its decision making mechanisms, but they are evolved models informed by the data they were trained with, so there is no explicit knowledge or reasoning processes to interpret or evaluate. The same concerns apply for existing comparison metrics for high dimensional generative models that are either qualitative or are cover metrics that are more suited for evaluation than for comparison. What is needed is a flexible, open-source comparison framework that can estimate the relative performance of existing systems. Benchmarking is an essential part of the scientific method, as it allows to compare the performance of different researchers and teams with the same problem (T. H. Menzies et al., 2021).

5. Open-Ended Learning and Artificial Consciousness

Stuart Russell recently proposed the following question concerning the future of AI and more generally, to what extent should we attempt to build an intelligent, autonomous system which could autonomously become the person who makes all the important decisions in our lives? Russell's short answer is that the AI should not become a person in charge. We may say that, at least for now, we don't want AI to reach a form of artificial consciousness. The purpose of our chapter is to discuss open-ended learning in the light of some important hypotheses concerning the nature of human consciousness. We will argue here that conceptually, open-ended learning is a necessary adjustment to the human mode of learning.

5.1. Concept of Open-Ended Learning

In some sense or other, open-ended learning has been researched since the very beginning of artificial intelligence research. Open-ended learning is the process through which adult people learn new things. Rather recently, a large community has emerged conducting experiments with deep belief networks in order to produce new works of art such as paintings or music. In addition to art, a new, more difficult domain still is the domain of self-play in complex games, unsupervised by conventional metrics as winning and losing. However interesting these developments may be, they appear to be far from what we mean when we say "open-ended learning." As of today, no AI is capable of making new discoveries, be they in science or elsewhere. Furthermore, in addition to art, a new, more difficult domain is the problem of text generation, unsupervised by conventional metrics, i.e., which has nothing to do with winning or losing. What we mean by open-ended learning is a capacity to acquire new representations of the world.

A notion that is conceptually related to the one of open-ended learning is transfer learning, the capacity to learn about new domains, related to previously acquired skills.

5.2. Theories of Consciousness

What do we know about open-ended learning in humans? Let us recapitulate some of its main features taken from the theory of consciousness. A first remark about human consciousness is that it is stubbornly private. We suggest to be quite sure that other persons have qualia similar to ours and also to believe that there are certain similarities, but from the fact that you are reading this paper, I cannot help but believe that I am the only being to have developed qualia. A second remark concerns the way in which consciousness appears to children. The world appears to them first as a collection of objects, then as by-products of actions. The claim that I will support is that the consciousness of action by-products comes first, and the consciousness of objects develops later.

5.1. Concept of Open-Ended Learning

Introduction to open-ended learning, principle of open-ended learning, Examples of open-ended learning, Some implications of the open-ended learning principle, Related concepts of progressive learning and lifelong learning.

By open-ended learning, we denote a system which evolves to become more complex and sophisticated with experience on its own, such that its capabilities are not predetermined or exhaustively specified ahead of time. Animals and natural intelligence show this ability in very obvious ways. Self-biological evolution exhibits open-ended learning principles, creating slowly by means of reproduction and natural selection new species with new abilities that did not exist before, including deep intelligence.

Another typical example of such processes, on a much smaller time-scale, is the learning of a child, or the social learning of a small group of people. Any system capable of creating, in the right conditions, more intelligent agents by means of interactive learning, is an instance of open-ended learning on a smaller time-scale, since the system itself creates the open-ended learning conditions and the dynamic of the process. Continual increases of intelligent capabilities have been observed in humans of groups with specific isolated cultural traits, during long time-spans of many thousands of years.

But there is also a smaller miniature embodiment of open-ended learning principles, in the natural world, which is the early development of a newborn in the first two years. Here we have one single individual gradually increasing its intelligence, through social and interactive experience. This seems to be the most optimized and efficient system to achieve high intelligence levels in those species with social groups, because of its low energetic cost (J. Togelius, 2024).

5.2. Theories of Consciousness

This definition of consciousness is too abstract to be useful, so it requires a more concrete definition based on more thorough theories. Several attempts have been made to define

and thus clarify the meaning of consciousness, its implications and relevance. There are two broad categories of theories of consciousness, introspective theories and philosophical theories. Introspective theories focus on the experience of consciousness from the first-person perspective, while philosophical theories focus on the study of objective and factual explanations of the consciousness phenomenon. First, we review both categories and what they consist of.

Introspective theories describe what we feel when we consciously perceive things, what it is like to witness them and apprehend them. We say that we are conscious of things thanks to a faculty called attention. Conscious perception is not automatic; it requires a certain effort that implies the involvement of our cognitive mechanisms. We also feel as if there is a special aspect of our experience, a special quality we call qualia. While independent of attention, qualia are intimately related to consciousness. The most important is the experience of "what it is like" to feel sensations such as the automatic bodily actions of reflex sensors, or taste sensations from sweet or bitter tasty fruit. Qualia guard a unique and special encapsulation of our experience. Given that the inside-outside structure seems to be enough to account for both qualia and attention, we can use it to describe the workspace mechanism, and thus define the Zoomers specifically in terms of qualia and attention (S. Dehaene, 2014).

5.3. Implications for AI Development

We believe there are at least two important implications that these paths would have for the development of AI. The first concerns Open-Ended Learning (OEL). Although OEL has been, up to now, a feature of low-level learning in the form of reinforcement learning, its fullest expressiveness would be achieved by enabling OEL and high-level learning to operate in partnership throughout the AI agent's life and for each other's mutual benefit. In a healthy life, the former would present the latter with an ever-expanding vista of environments, problems and tasks to learn, and the latter would use its model of the world to come up with new information, knowledge and skills, suited to the agent's specific experiences and exploration goals. Just as, in humans, most of the knowledge accumulated in life comes from external sources, but the most influential knowledge comes from internal reflection, we believe the greatest dividends will come from partnerships where OEL is primarily responsible for feed-forward, experience-driven knowledge generation and high-level learning is primarily responsible for processing that knowledge.

The second concern is the AI's relationship with the world, its subjects and the subjects' minds and related processes. As we said before, partly originating with Piaget, OEL is not a mere adjunct of high-level learning. On the contrary, it is the indispensable, enabling background of high-level learning. A high-level learning agent needs its OEL system fairly well aligned with its design goals. Such alignment is a necessary, albeit not sufficient, condition for high-level learning to result in useful knowledge. It is therefore essential, not merely desirable, for AIs using high-level learning to have minds where OEL accomplishes developmental and control functions similar to those in biological

minds and to those essentially specified in cognitive developmental theory (T. Schmiedl, 2021).

6. Conclusion

In contemporary times, we have achieved remarkable milestones in cognitive AI. There are inspiring demonstrations of systems that use cognitive science concepts to make strides toward truly generalizable learning outside of the training data. There are cognitive systems tested on provision of truly wide-domain language use. We understand language best when it is used to act on the world. Systems that allow the language instruction of embodied systems are coming close to fulfilling the proposal of "language as chariot". Progress is being made toward actionable agents that can think throughout the process of planning and that can learn to plan by doing. Embodied systems with good "popular commonsense" are learning language by interpreting the intentions of the agent that they are trying to imitate. They learn the meaning of words by interacting with the world. In an era of vast neural models performing huge learned tasks, they are the exception. However, still our conceptualization of cognitive AI remains a rusty toolbox. AI is driven by tasks and demonstrations of progress on benchmarked tasks. These task-driven benchmarks are unrealistically limiting, and how the typical task will culminate in the completion of the goal is mostly not specified. This is also true for embodied systems. The considered choices of very big models trained on the entire corpus of language are constructively valid choices of cognitive AI.

As we seek to scale to the grand challenges of cognitive AI such as common sense understanding, multi-task and zero-shot generalization, and explanation and planning, we still need to focus more on understanding the intimate patterns of neural computation that truly realize generalization. Cognitive AI seeks to understand how the brain accomplishes the remarkable feats that it accomplishes. There is not enough emphasis on testing specific guiding theories about these intimate workings of cognitive AI. As an outlined area, we are still spaced along the frontiers of cognitive operation in cognitive AI. In any case, as the term AI increasingly incorporates cognitive modeling, the searchlight of neural models will throw a shadow on the how-questions of computational neuroscience.

References

Ahmad, A., & Gulliver, T. A. (2024). Green AI: Sustainable and energy-efficient artificial intelligence. Springer.

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

Danaher, P. G. (2021). Automation and utopia: Human flourishing in a world without work. MIT Press.

Dehaene, S. (2014). Consciousness and the brain: Deciphering how the brain codes our thoughts. Viking.
Ghosh, A., Ghosh, B., & Hassan, M. D. (2023). Neuromorphic computing and beyond: From algorithms to hardware (1st ed.). Springer.
Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust (1st ed.). Pantheon Books.
Menzies, T. H., et al. (2021). Practical artificial intelligence: An enterprise playbook. O'Reilly Media.
Molnar, C. (2023). Interpretable machine learning (3rd ed.). Leanpub.
Narayanan, A. (2023). Privacy in the age of AI: Training models, protecting data. Princeton University Press.
Narayanan, A., & Kapoor, S. (2024). AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference. Penguin Random House.
Patterson, S., et al. (2022). Carbon emissions and large neural network training. AI Index Report.
Schmiedl, T. (2021). Meta-learning: Bridging the gap between learning and experience. Springer.
Shams, B. L., & Farhangi, A. (2021). Explainable AI in healthcare and finance. Cambridge University Press.
Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. Alfred A. Knopf.
Togelius, J. (2024). Artificial general intelligence. MIT Press.
Wateler, S. M., & Kaiser, M. (2024). Ethical trust-by-design in AI: A framework for dynamic human–AI partnership. AI Ethics Review, 4, 101–120.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-611-9

DeepScience
Open Access Books

Chapter 5: Navigating the Future of Cognitive Artificial Intelligence

1. The Future of Cognitive AI: Research Directions and Breakthroughs

Research Directions and Breakthroughs: Enhancing Foundations with New Ideas

Cognitive AI research creates cognitive breakthroughs. The typical focus has been on scale and architecture with secondary attention towards better understanding and exploiting "what's happening" inside the best models. Efforts to better document these efforts include gathering resources on cognitive NLP, cognitive vision, and cognitive speech. Lately, there have also been efforts to better estimate realistic long-term societal impacts by measuring intellectual peer effects using data from exam season. At the same time, the industry is consolidating; industry experts are working in new startups which are creating cognitive Democratization tools.

Research directions are clustered around current hot areas of focus. On the Cognitive Tools side, a large majority of papers are devoted towards better understanding the few-shot NLU and the tools that work across modalities including vision, language, and speech. These tools are released by major technology companies, and increasingly the models are being improved via system cards. The second direction improves modular solutions and probing. Modular probabilistic reasoning solutions work particularly well in creating decision pipelines for complex applications drawing on multiple experts. Probing seeks to understand model failures by better understanding sample quality requirements, or better understanding neurons specialized for specific tasks. Other research improves human-centered approaches. By incorporating delays, one kind of research builds Cognitive Agents which optimally combine human effort/input and Cognition-AI. A second kind observes cognitive effort itself, perhaps by better engaging with mishaps and generally creating model cards for better usage.

A major long-standing direction is creating cognitive ethical assistant machines. These efforts include both determining when it's a good idea to send cognitive effort to humans (or hybrid assistants) or automating decisions for when to leave cognitive duty to AI, integrating explanation and transparent model design whenever possible.

1.1. Current Trends in Cognitive AI Research

The modern Artificial Intelligence (AI) era is relatively short, with the launch of the modern AI in 2012, when the first neural networks were used to detect images at a level better than humans. This led to a series of revolutions throughout the data-driven world whose applications have been increasing rapidly and exponentially into areas of human physical capabilities; microphone tasks are now non-speech processes; vision tasks have turned into imposters of the human eye capabilities; robot control takes the guiding into levels of real-life dexterity; analyzing summarizing, interpreting, and even generating text and other sequential tasks are being executed by transformers in laughable paces. But are AI solving the tricky tasks of understanding? The answer is no. Many will argue that this is a separating wall, whereby the extremely good statistical pattern matching achieved with deep learning, fine-tuned on extremely large semantic labeled data, should be hyperbolized. Symbols that humans explicitly learn and use should become implicit ingredients in machines taking over cognitive tasks, and the influence of the semiotics symbolism interpretations of the human mind should somehow be transferred to AI. But there is no doubt that AI is currently far from employing meaning and that large language models generate hallucinations that stress the nature of the statistical modeling.

In this context – known as cognitive AI, the AI that flirts with understanding and uses representations and knowledge – there are a number of questions still open, leading the cognitive neurosciences investigation as well as philosophy sciences on our innate derived and perceptive capabilities. For instance, can superb auditory-graphic symbolic models discover common-sense knowledge in the form of Patchy facts? The majority of generative models are completely empty networks from an internal representation point of view, but should indeed then be symbol manipulating systems, able to create dynamic records of event patterns in time and space, constructing meaning-based mental representations, for the cognitive tasks of understanding, reasoning and common-sense knowledge (C. Summerfield, 2025).

1.2. Key Breakthroughs Shaping the Future

Recent breakthroughs in deep learning, such as reinforcement learning from human feedback, diffusion models, multimodal capabilities, long-context models, and alignment of foundation models with human values, have significantly increased the usefulness of AI systems in a variety of applications. I argue that the next decade of Cognitive AI research will focus on these five themes. Specifically, the ability of Cognitive AI systems to learn from and talk to us, their creative skills in generating novel content across data types, the integration of multiple modalities within and across tasks, their retrieval and reasoning capabilities over long bodies of text and data, and their ethical alignment with our human values. Generating artificial sounds, images, and text documents has a long research history, going as far back as 30 years. However, the recent increase in demand for novelty and content creation, spurred on by the incredible performance of foundation generative models, have revived focus on this line of Cognitive AI research. Unlike the practical use of generative reasoning in many traditional sense, such as computer graphics, style-transfer, and data augmentation, the

generative capabilities of foundation models address the need for AI systems that can help or do us creation of new content. This push of creating content from wider audiences can have significant benefits. We can create new form of media that can be cheaper to generate, in areas such as video production and game design (J. Zou et al., 2024)

1.3. Ethical Considerations in Cognitive AI

This chapter briefly touches upon an ethical consideration: that AI and machine learning programs display bias towards races that have not been favored or highlighted in the datasets given to the programs. Although some may think that this chapter's inclusion of these topics is anti-AI and more of a sore thumb, it is critical to discuss the ethical considerations of a technology that has been increasingly becoming popular, and now governs decision-making for various sensitive areas like college admission, hiring, and loan applications.

Machine learning algorithms use heuristics — these are rules or principles that lead to solutions with a reasonable speed. Such programs can recommend to a clinical psychologist if a patient should be recommended for a drug-based therapy based on the patient's condition, previous histories, choice of therapy, and even scrutinize things like microbiome composition in order to determine what drug-based therapy would best work for the patient. These algorithms are driven by scores of professionals who mostly develop and deploy these kinds of systems: Are they trained to build these systems in an ethical manner, one that does not disadvantage certain patients or populations? Some of the answers lie in auditing how well these systems perform structurally for different ethnic and social classes. However, are domain experts and technical experts actively working in tandem to assure that the algorithm is fair to all classes? It may be observed that given the burden of regulatory fatigue and documentation costs, most such systems are predominantly deployed and monitored by professionals (J. Buolamwini et al., 2024)

2. From Deep Narrow to Broad General: Paradigm Shift in AI Thinking

The recent breakthrough in generative AI treatment is built on style models. However, the models on which this core achievement is built are still Deep Narrow: optimized around a single or minimal number of objectives and employed primarily in closed loop settings. They do not optimize for multimodal performance across heterogeneous data and tasks, nor can they do so without additional fine-tuning and adaptation. Furthermore, they are not estimated to be the only best performing model for the skill. Nor is that necessarily a bad thing. Building permanently-in-top-working-order multiple-peaked models is expensive and involves hard engineering trade-offs. Even so, in the long term, the interesting real world technology is moved by Broad General AI, and doing so requires a transition to Broad General models that can potentially simultaneously operate across all or most cognitive skills multimodally within one framework. This capability exists in many real world humans, and the Broad General models, as well as tasks and optimization, may provide additional helpful insights or approximations.

The move to Broad General models will bring a few key changes. The core task formulation will shift itself from low level prediction tasks of priors of particular cognitive skills to having the multi-prior multimodal functionality for the entire cognitive ecosystem as a central design goal. A set of diverse auxiliary parametrized or non-parametrized guidance techniques are employed to guide atypical tasks. While auxiliary guidance may also be mixed with task-based fine-tuning for specific tasks in a transfer scenario, key trade-offs about what auxiliary techniques are needed for a skill or skills, task durations, and usage storage space etc, are modeled and optimized by the Broad General models themselves in an transition. Note that this scenario talks about the Broad General model being a multi-skill multitask system, having large memory capacity.

2.1. Understanding Deep Narrow AI

Deep Narrow AI (DNAI) is the embodiment of narrow AI systems that focus on single or limited narrow tasks based on vast amounts of structured and unstructured data. These high-performing models, such as generative AI natural language models and image recognition, exceed human performance. Such models go significantly beyond conventional narrow AI rule-based models while being deeper with billions of connected nodes learning from vast datasets. These systems have ushered in an unprecedented shift in natural human-machine collaboration; many of the deep AI model systems have become indispensable tools for common and specialized tasks. Many white-collar jobs, ranging from lawyers creating and reviewing contracts to startup founders writing business plans and software developers coding programs, have increasingly relied on these models to save time and improve performance. Despite its awesome capabilities, DNAI lacks true human cognition and creativity, which is limited to the narrow tasks it was designed and taught to do.

Deep AI models are narrow because humans are incapable of training AI systems with the vast amounts of input data needed to imbue DNAI with broad general human cognition of the world and creative AI capabilities, including the capacity to learn without human intervention. Moreover, traditional AI-wide human-predictable systems operate on the principle of programming specific features of tasks to follow instructions, heuristics, and rules to be done. In contrast, deep narrow AI systems are complex and sophisticated, drawing inferences, associations, and conclusions to decide a course of action. For example, a model deftly composes a business plan, including producing a pitch deck with prompts about the model's capabilities, company description, products and services, market analysis, marketing strategy, competitive analysis, and management structure. Thereafter, it describes each departmental duty in detail (Y. LeCun, 2024).

2.2. Transitioning to Broad General AI

The broader transition toward Broad General would require building innovative and greatly expanded phases, systems, architectures, and strategies, that enable a smooth, efficient, dependable transition from Deep Narrow to Broad General. The initial Broad

General phase would enable substantial performance improvements on many capabilities like Deep Narrow, but fundamentally shift the paradigm in how we build and deploy Broad General, including allowing greater model dilutions that directly improve impact per dollar. We expect the first Broad General systems to be substantial modifications of existing Deep Narrow systems; however, with strategic thinking and an appropriate research push, it is possible to lay down the novel foundations of radically differently architected Broad General systems. Later Broad General phases would enable Broad General systems with significantly larger model capacity but at greatly reduced per-capacity inference and training costs. Finally, advanced Broad General capable systems would radically shift both capabilities as they move toward AGI, and capabilities-cost scaling—affordable Broad General AGIs inside the homes and communities of hundreds of millions. Indeed advanced Broad General may open up novel pathways to considerable capability progress in widely diverse areas of human preference/maximization such as science and engineering, art, design, health, medicine, etc. owing to its input-output modes that rank very high along the two axes of personalization: diversity of outputs to maximize richness in life experiences, and focus on life effects for a given relative weight on self vs. community effects (R. Kurzweil., 2024)

2.3. Implications of the Paradigm Shift

The paradigm shift being undertaken in our thinking about AI allows us to specialize specific aspects of cognition while trusting in the ability of the AGI to integrate these aspects into a coherent whole. Cognitive AI provides a single framework for addressing the entire spectrum of capabilities required for human-level artificial intelligence, such as discussion, planning, decision making, language translation, inference, generalization, coordination, and knowledge management. This unified understanding opens a path for parallel and independent development of all cognitive capabilities, while also enabling seamless integration into a cohesive whole. This is unlike narrow AI where no matter how successful an individual capability becomes, it still acts under the invisible constraint of being just a specialist, never able to assume the mantle of being general. Cognitive architectures provide the ability to progressively assemble any specialization of a cognitive capability as an agent moves from infancy towards humanlike intelligence. Cognitive AI architected with a strong sense of autonomy and awareness can then not only manifest, but also create and question models of the world, coordinate different capabilities in support of a larger goal, exercise value judgments as situational context evolves, and above all else, act upon the world. This is the hallmark of adult intelligence, and the guiding goal of Cognitive AI systems (R. Kurzweil, 2024).

3. Global Collaboration for AGI: Academia, Industry, and Governance

In the push toward Artificial General Intelligence (AGI), no single organization can take on the challenge alone. While collaborative models demonstrate that tightly defined

partnerships can enable rapid productization, we need to think about deeper partnerships and loosely defined cooperation at a much more systemic level. Here, we propose a model that brings together industry, academia, and governance to take on this monumental task, whether it is to hire the best people in each sector or truly understand the ethical challenges and threats posed by AGI.

In the early years of AI development, considerable effort came from academia, where researchers were able to intellectually explore the deepest theoretical questions of intelligence. Many still dream that we can create AGI as the by-product of some theoretical insight. Research in AI safety and alignment is still strongly grounded in the academic world. However, funded at an accelerating pace by competitive investments, AI has gone through an enormous transition into a commercial product—with some dramatic early successes. Today, large language models are democratizing content creation—both in written media and by extension into many other modalities through multimodal models. Developments are trading speed for thoroughness and safety, with live experiments unfolding in the wild.

AI has entered a self-reinforcing cycle of public success—demonstrating impressive capabilities—and private investment—competing with hyperscale investments more generally. However, proponents of AGI, and of ethical guidelines around its development and deployment, consider the industry–government interface at this point to be sorely underdeveloped. The ethical and societal challenges surrounding an exploding capability class are left principally to private companies—often with different incentives than those of the ethicists advocating a careful, systematic approach. A maturity in global governance regarding data privacy, disinformation, recommendation-driven behavior modification, the risk of generalization beyond training, or of the creation of invisible internships in content moderation also has not kept pace with the increasing speed of industry.

3.1. The Role of Academia in AGI Development

The quest for AGI should be open-minded, and perhaps the only way of avoiding the plethora of vile AI pits is by allowing the wit of a myriad of curious minds to collaborate. But, while the University exists for knowledge in and of itself, industry funds cleverness having declared intent to embroil AGI development within a cocoon of, what some might consider, purposeful ignorance. Government acts as overseers of societal demands, and so oversees also the custodianship of the ASI technology. It should also look for ways to catalyze collaborative musing, but how? What role should each actor play? Universities will, like any other entity, lean towards profit, convenience, and power. Corporate-funded AGI development, however, is subject to monopoly, and sudden bursts of genius might get stuck in the folds of time due to corporate misaligned goals. Extreme caution needs to be exercised concerning the puppeteering strings held. The role of Universities would be to be the talking points of difficult problems. Look deeper into the impact of technology rather than solely the potential results. Seek odd-looking corners to confirm or rule out. Search for the shaft of light within the AGI black box of

corporate partnerships. Be wary of intimidation and habit. Society craves answers. The role of the academic pillar is that of trusted caretaker. But is it solely that of a caretaker looking after a domain of unknowns while being held at bay by capital gained impetus? Although Universities act as custodians, they might also have an important role in ushering in the AGI breakthrough sooner than would normally happen. The interpretations presented would basically be fuelled by curiosity rather than business directive. Funded power explorations presenting counter-arguments to prevailing theories, and outside interpretations delving deeper into sociological, philosophical, or even ethical facets (A. Narayanan et al., 2024).

3.2. Industry Innovations and Partnerships

In coining the term "cognitive AI", we intend to keep the focus on Generalized AI and its Cognitive capabilities. However, companies are also exploring its application-specific behaviors, such as geolocation-based assistants, automation products, or other implementation-focused, architecture-specific niche products. The progress from zero to more restricted application-specific capabilities is absolutely amazing. But the larger dialogue should still center around Generalized AI and its Cognitive capabilities.

Without minimizing any of the above innovation, we believe discussions would serve our industry and our world better by grounding them in the context of Generalized Cognitive AI and its capabilities and by establishing a clear set of short- and long-term benchmarks around the space for both industry and society. The scientific exploration and the underlying research agenda could be grounded and driven by academia, while the short-term innovation would be for industry to pursue. Partnerships would be a natural extension, but the roles would have to be drawn clearly. Finally, public gaze and the nature of our society could shape both the academic and the industry efforts. It would take all three to reach the optimal AGI state.

In that spirit, it was quite encouraging that even in the early days of the current bound of capabilities seen in industry innovations, many of our leading universities have modified their specific Computer Science programs and established new Cognitive AI or AGI specific areas of interest and research. Industry too is reacting and evolving upwards. New offerings and partnerships are also taking place. One company is leading with its significant investment and partnership around a major AI product. Other players too are forming partnerships and investing. How this shapes over the next long while is still an open question (A. Narayanan et al., 2024).

3.3. Governance Challenges and Opportunities

All the governance frameworks we have today are in need of critical updates in light of new eras of AI development and transition towards more digital, data-driven systems. The acceleration of technology development, especially in light of the recently raised pathways to more generalizable, robust, and capable AI, necessitate the need for review, change, and established response systems for anticipated (and unanticipated) by design and consequence risks. We are in unprecedented times when technical innovations and

developments create new openings for reshaping old problems, as well as enabling the flourishing of new – and potentially even more harmful ones. Through this period, there is potential for increased stakeholder mobilization and partnership to identify, re-evaluate, and reshape the categories through which we assess the systems we build, the responsibility of those systems for flourishing sustainable and equitable pathways for development, and the shape of governance interventions and incentives in a way that not only responds to possible new risks and safeguards from harm, but also shapes and unlocks futures and use-cases where AI technologies feed into human flourishing trajectories. Technologists are often the first to notice changes and horizon-scan for opportunities and challenges on computers – we have already seen discussions on the risks surrounding the production of new types of persuasive or misleading content through computer vision AI-generative systems. However, the opportunities that are malleable with the roll-out and opportunity possibilities which are ushered in by the technology should be evaluated as well (A. Narayanan et al., 2024).

4. Your Role in the Cognitive AI Revolution

While Cognitive AI may generate images, text, art, music, and video, you will be the curator of it all. You will gain expertise in your field or industry and be able to use your imagination and creativity to bring together ideas and concepts in a manner even the most sophisticated of AI cannot. You may own the domain of your expertise fully and dedicate yourself to it and employ Cognitive AI as your assistant to help you with the small things, or even dedicate a portion of your time doing so while also using Cognitive AI for other projects. You may freelance for several small to medium businesses in an expert area or assist professionals with their work—just as you do now. The role of AI you require for your business and future may be quite different than the role of AI that transforms work for the average person in a corporation or medium-to-large-size business.

Understanding Your Impact

Currently, companies employ Cognitive AI for customer interaction and service, data processing, foundational user research, drafting copy and writing, language translation, programming, scheduling, simulations, testing, and training. Digital creators use Cognitive AI to generate images, video, and text. The Cognitive AI expense for company processes will likely go from millions of dollars a year to just a few thousand, and its use will skyrocket. The more companies get used to existing with Cognitive AI as a vital part of their processes, the more new companies relying on it will get created. As with past technological transformations, attendance in the AI-related job market will likely spike and peak once Cognitive AI is integrated into most people's daily devices or processes. Then, much corporate and individual work will eventually focus on utilizing the new AI tool and skills.

Learning how to evaluate, work with, iterate on, and improve AI is going to be a major driving force in business. Any nature of a task that requires minimal human input, time, energy, and guidance will be replaced. Cognitive AI will eventually become another type of vendor. Understanding that, as an individual or business owner, will help guide your experience with it. The need for human oversight will strengthen because humans are more adaptable in more nuanced and unpredictable situations, and there are more physical tasks performed by a physical human.

4.1. Understanding Your Impact

As cognitive AIs become increasingly integrated into societal frameworks, workers will begin to feel the impact, particularly with Large Language Models. A stable internet connection and a device with the ability to connect will enable anyone, anywhere to generate written content. From full articles to personal letters to creative fiction and poetry, nothing is beyond the reach of generative AIs. Coupled with other models and asset generators, there truly is no limit to the generation. In this future, the generational abilities of these systems will remove the need for humans to generate any of the work effort or knowledge - but can you imagine a world where nobody needs to use their brain?

The undertaking of generative work is what separates us from the animals: cognition is hard, tiring work. Creative pursuits are also a key unifier of our humanity. Removing the labor demand from this generative work, particularly in areas such as fiction or art, places many creative pursuits into the realm of hobby. For those who make a living writing books or crafting art, these systems offer potential disappointment and disruption. Will customers opt for cheaper, generated books or art pieces instead of your master work? Will your works still hold value? Will you be able to eat if your entire field becomes oversaturated with content and understanding the delicate balance between the value we place on creator work and the abundance on offer becomes impossible? It is highly likely that generating tasks will no longer be monetizable, and that the unique selling point for those who do not embrace hobbies or humanity would be invented, creative work – the task that at present is quite hard for AIs to accomplish (A. Lee, 2023).

4.2. Skills for the Future Workforce

While some occupations may be displaced by the advancements in automation and AI, newly created jobs will require a specific mix of capabilities and higher-level skills. Prior research indicates that the future workforce will rely on three main skills to work with AI and Cognitive Technologies: knowledge, creativity, and interpersonal skills. While, traditionally, knowledge work focused on tolerable task automations improving the efficiency, speed, or accuracy of workforce processes, assisting augmented robotics and cognitive AI solutions will be most successful when enabling the workforce to bring more value to the organization by emphasizing either creative or interpersonal capabilities associated with socializing and developing ideas or engaging other humans.

Engaging AI-supported enterprise resource systems will put organizations in a better position to reduce errors or streamline workflows but may not necessarily provide higher-quality solutions for producers, customers, or partners. Successfully developing the collaborative ecosystem, integrating intelligent tools into their work, and embracing AI will determine the future success of the organizational workforce. To face the creativity and interpersonal needs of the new task environments, organizations will need to nurture and develop their workforces' Creative Intelligence: the knowledge and ability to effectively deliver the key interpersonal and introspective builders of originality, critical thinking, and idea development; curiosity, questing and structured intelligence; concern, empathy and communication skills; collaboration, teambuilding and active listening; and courage – the willingness to act on one's ideas. To reskill upward, the future workforce must seek out programs designed to help them develop the affinities defining innate capacity for success in these interpersonal and creative skills. Successful programs addressing these needs typically focus on simulated collaborative environments for creativity, courage, empathy, and communication. Programs focused on inquiry, engagement, ideas, and initiative can feed creativity and curiosity. Programs developing original and critical thinking and structured exploration can bolster structured intelligence and questing (R. Susskind et al., 2022)

4.3. Engagement in AI Communities

Platforms that allow for clear and mutual communication are at the center of how a community can engage. To mention a few examples, some platforms have gained a lot of traction due to their straightforward usability. One platform is a social news aggregation, web content rating, and discussion website that enables users to submit content such as text, links, and video posts to the site. The submitted content is then voted up or down by other users, and the most voted content moves to the top of the user-submitted listings. Later, it reached a staggering traffic of 1.2 billion monthly visits. Another platform gained attention for enabling free voice, video, and text messaging, and channel organization for groups. Originally designed for video game players, it now hosts open groups of people with shared interests. Established scientific communities now collect a large quantity of 21st-century research. In fact, to engage, we will need a more coherent and unified approach to identity in these communities, considering super-intelligent AIs and deceptively persuasive tools.

The goal of this section is twofold: to present a relatively small fraction of AI-related communities that already exist; to encourage readers to join these present and future communities. The motivation behind this is simple: these communities will shape the future of cognitive AI, and thus the future of humanity. For their work, volunteers in these communities often go above and beyond what is expected of them. Their incentives will help boost a general sense of purpose, community, and fulfillment within membership. This sense of belonging, and commitment to a cause greater than oneself will, naturally, lead followers to avoid or criticize behavior that threatens group integrity (R. O'Gieblyn, 2021).

5. Challenges in Cognitive AI Implementation

The promise of cognitive AI obscures an inconvenient truth: the path to implementation is littered with obstacles. In a technology space already recognized for its disruptive capabilities, Cognitive AI stands apart in one important sense: the difficulties of implementation stem not only from known technological hurdles but from questions of intent, trust, and the fuzzy line between benevolent enhancement and pernicious replacement. These difficulties are magnified at scale, both because of the impacts that scaling may have in terms of user experiences and redesign, data and model transfer biases, and training heuristics, and because of the social impact of amplifying bad decisions made by individuals or small groups. In many regards, these challenges are not fundamentally different from those faced by any new technology that quickly crosses operational boundaries to scale: widespread adoption of facial recognition, for example, has generated backlash and proposed regulation due to negative impacts among certain demographic groups and the chilling effect it has had on speech.

But Cognitive AI differs from either of these examples in its potential to serve as a new point of contact between users and information. Trust and misinformation therefore take center stage, on the one hand because of the demands Cognitive AI will place on the existing information ecosystem and on the other because of the fact that users' willingness to interact with Cognitive AIs can significantly affect their market and operational models. The potential consequences will be more profound and widespread at scale, meaning that the solution-space of institutional adaption is larger as well.

Many hurdles and concerns may be qualified or subcategorized under this umbrella. For this section, we will broadly outline the three main areas seen as leading to a high barrier-to-entry for these technologies: technical barriers regarding the practical optimism around the application of Cognitive AI, resistance from society, and finally agreements on regulations and legislations regarding AI use.

5.1. Technical Barriers

With the increasing popularity and availability of Cognitive AI technologies, many organizations are exposed to the latest developments, products, and services in this area. Still, the transition to real-life Commercial Cognitive AI services is difficult and has several challenges. As such, concerns about the technical barriers preventing company executives from adopting Cognitive AI tools are being raised. Our knowledge and experience within the Cognitive AI domain and specially about the Cognitive AI industry help us to better understand these existing barriers. In this chapter, we analyze these existing and probable future barriers in order to answer our research questions and help others in meeting these challenges. In addition to the pace, breadth, and depth of Cognitive AI technologies and their adoption, many existing and possible future technical barriers exist. These include the complexity of business design requirements for the development of new services that are based on Cognitive AI technologies, a lack or inadequate availability of adequate Cognitive AI tools for general use by the service

industry, the quality and availability of Intellectual Property Rights for Cognitive AI developments, high development costs for new services requiring tools developed in-house, high expenses needed for utilizing third-party Cognitive AI tools, and a lack of appropriate methods for the evaluation and comparison of Cognitive AI services. Overcoming only some of the barriers may be sufficient for recognizing short-term opportunities to generate a variety of Cognitive AI services. If so, the recommendations presented in the text provide an easy way to exploit and recognize some of these opportunities. Companies and other organizations not only increasingly recognize and monitor the fast-paced developments of the Cognitive AI domain, technology risks and gaps, associated with the adoption of commercial services, but are also perceiving business opportunities to create new services or improve existing services. To answer the question of why there are existing barriers and risks to the adoption of more popular commercial Cognitive AI services in the short term – and to facilitate access to those opportunities – we also discuss the other possible barriers, such as personnel resources and expertise, and internal company management and organizational structure (M. Iansiti et al., 2020)

5.2. Societal Resistance

In recent years, the Great Resignation has taught us that many employees are willing to forfeit higher wages in return for other value...such as job satisfaction. New technology is one of the clearest signs of innovation, creativity, and job growth. But if, in their heart of hearts, employees don't want to change their workplace habits, they might push back against innovation that would otherwise create new avenues of revenue and purpose for them and their companies—especially if those innovative technology tools are cognitive AI applications that often erode career-building tasks, replacing internship positions with machine learning models and human oversight. Those cognitive applications that augment what humans do for a living tend to perform better in terms of ROI than those that replace human efforts. Along with company-sponsored training and development programs, the potential friction of labor manipulation can be addressed by a focus on augmenting existing jobs rather than eliminating them. Cognitive AI advocates and executive leaders should tune into the clamor for new roles and new responsibilities, vet the angry tweets that question the validity of a role that has been impacted, and speak at internal functions about the value cognitive AI brings—its ability to promote creativity and upskill workers.

Of all the challenges to adopting cognitive AI technologies, resistance from the frontline, from employees who care deeply about their positions and practices, will be perhaps the hardest challenge to overcome. Those positions—the mundane roles that suck up available time, resources, and budgets as deadlines loom and armies of employees input the same repetitive data tasks to get a special project out the door—are the gilded cages that are the emotional and financial lifeblood of any company. Inside those gilded cages are the tantalizing, announced potential of life-affirming, fulfilling positions that keep companies alive and in the black. Yet, even inside those glittering cages, employees still need massages and beer Fridays, sports, and above all else, purpose (D. Susskind, 2020).

5.3. Regulatory Hurdles

Organizations are incurring the risks of using or integrating Cognitive AIs without sufficient knowledge of how regulations affect liability, privacy and data governance regulations, and enforcement of contract violations. We have already seen the curtailment of affairs of organizations in various jurisdictions as a response to the regulatory frameworks. The newness and the rapidly iterative nature of Cognitive AIs mean that there is vibration across the regulatory structures in their particularity. Such particularities exist in intent and enforcement at times in areas such as copyright, unfair competition, patent infringement, trade confidentiality, brand trademark, and product liability. Older nations or jurisdictions have begun efforts at either specific AI legislation or more likely collaborative actions across their borders. The traditionally quiescent GDPR recently promulgated a framework that, while generative AI-non-specific, does have reverberations across the compliance teams of organizations that have already incorporated Cognitive AI Technologies or are in the mix to do so.

Not only are jurisdictional borders disanalogous in terms of enacting and regulating, the uneven nature of the location capabilities of cloud organizations provides unique compliance challenges. The compliance teams in organizations may or may not know where the actual storage locations are. In addition, the particularities of the regulations over a period of time may need incursion upon internal approval processes to ensure compliance or to augment the processes to aid the execution of safe processes. Cloud infrastructure organizations also have the challenge of stating in easily consumable terms how the storage and processing and even the data input channels are secured against either repurposing or regurgitation (K. W. Abbott et al., 2024)

6. Future Applications of Cognitive AI

Where we are heading with Cognitive AI must be approached cautiously. Cognitive AI presents society with both a boon and a challenge. As we embark on the quest for increased intelligent, integrated, connected, and automated systems, we must also remain aware of all the trappings and pitfalls inherent in significant technological advances. The future we can envision, to a degree, includes automated systems in Healthcare, Finance and Risk Management, and Education and Personalized Learning.

Healthcare Innovations

In healthcare, innovations already underway will allow new types of decision support. Better knowledge about treatment planning and why-factors for adherence analysis for chronic diseases. For example, generating on-the-fly answers to questions by patients from the domain knowledge that can be extracted from the combination of the public Medical Knowledge Graph, the link data universe, and clinical resources. Another application perhaps would be to generate personalized boring adherent messages for patients from the knowledge of the patient's psycho-socio-economic attributes,

demographic factors, clinical history, family history, adherence to prescribed drugs, mental disease history, current social position, and the information about the disease he/she is suffering.

Finance and Risk Management

In the risk management field, banks would be able to attract customers that are considered “dead in the water” today and give them a second chance by offering smaller loans than normal, with the customizable terms and conditions that will make repayments achievable. Such a personalization is only achievable by using information and data from multiple data stores at the same time and leveraging on-the-fly inference. It is also possible to better tailor insurance offers and banking services in general and anticipate claims in the case of natural disasters. In Asset Management, it will be possible for Wealth Managers to create real time, on-the-fly portfolios in case of sudden life-situation changes that are determined by any of the various database servers available for querying.

6.1. Healthcare Innovations

The healthcare sector stands on the brink of a veritable revolution, with cognitive AI solutions already expanding its capability base. These innovations harness both personal data – genetic and medical condition data – and social data such as genomics and social connectome. The latter set of data derives from an AI ecosystem that learns about members of a society from public and open data available on the networks and thus builds a model of each individual as a part of the network-ecosystem.

Health innovation has been restricted due to epistemic constraints stemming from traditional approaches lacking tools and methods required to assimilate what has largely become complex chaotic knowledge. The empirical laws and causal-mapping rules emerging from network dependency of connectome data give rise to sophisticated cognitively based predictive digital twins that are capable of accomplishing both diagnostics and prognosis of health and disease conditions.

Cognitive networks will allow cost-effective big data record keeping in terms of human connectomes, and mobile cognitive platforms will permit real-time monitoring and reversible nudging of all members of the ecosystem. Actual medical advice and provided healthcare can therefore be reoriented towards preventive – rather than interventional or assistive – modes and methods. This is the area of digitized precision or personalized medicine, in which the key workload changes from testing and diagnosis to easily available and inexpensive temporal sequencing of all relevant diagnostic probabilities (A. Min et al., 2024)

6.2. Finance and Risk Management

Cognitive AI is opening up exciting new possibilities in finance. Research scientists have used reinforcement learning techniques to develop algorithms to automatically trade futures contracts on multiple exchanges in daily or weekly trading modes. These

algorithms have been shown to track the best of humans in terms of profitability, with higher speed and discipline. Such autonomous active management will reduce fees that investors pay to the best traditional hedge funds. Cognitive AI systems, however, do not stop there. They can trade across multiple markets, on multiple time horizons, and can also incorporate probabilistic forecasts of future return and risk distributions for the relevant markets. Cognitive AI systems can make automated deployment and liquidation decisions to ensure that risk does not grow too high while the portfolio is being built.

Machine learning techniques, specifically supervised deep learning, have been applied to predict the probability of default on corporate bonds, one of the most intensely studied research areas in finance. The results have been promising. Probability of default is a crucial input to credit pricing and value-at-risk calculations and increasingly big data from social media are being utilized. Financial institutions are using data from social media among other alternative data sources as input to alternative models to complement and enhance their pricing and risk management analytics. Banks developed their credit risk management systems using insights derived from big data and social media that had been trained on corporate bonds origination data (S. Ruder, 2024).

6.3. Education and Personalized Learning

Advanced AI systems have demonstrated great promise as pedagogical agents in environments that support personalized learning. From intelligent tutoring systems to more directionless discovery-based learning, these agents have exhibited the capacity to recognize and respond to learning trajectory impediments by providing targeted hints, prompts, and feedback. Such guidance helps learners explore foundational concepts, apply knowledge, and reach deeper understanding and coherence, enhancing problem-solving processes. Moreover, these agents can tailor content and process help to the individual needs, goals, and preferences of learners, sparing them the consequences of learning by failure or through trial and error. Increasingly, agents can interface with learners in natural language, often bridging gaps and misunderstandings that would be almost impossible upon mere text that no human being would produce, hence adding the needed human touch. Moreover, the energized co-creation processes made possible by AI-catalyzed dialogue allow learners to move in and out of lead roles in the interactions, mirroring real-life collaborative innovation and creativity in diverse STEM and STEAM contexts. Because many learners now have access to a smartphone or similar device, capable of interfacing with advanced pedagogical agent technologies, there is growing potential for these platforms to support personalized learning across student and teacher populations at scale. The interaction paradigms and intelligent algorithms continue to advance rapidly, and the challenges that remain are primarily ones of innovation, inside and outside the classroom, in the design of engaging tasks, co-design of agent-supported inquiry processes in partnership with learners, teachers, and a range of content providers, and then effective study and evaluation of supporting evidence at diverse scales. Technology design and co-design, task design, and study and evaluation are key to realizing the vision of Learner-Centric AI and advancing the Science of Learning technology direction (B. P. Woolf et al, 2023).

7. The Role of Data in Cognitive AI

As the old adage goes, “garbage in, garbage out.” The effectiveness of AI models and tools in solving business problems depends largely on the amount and quality of data fed into those systems. For AI-based models or tools to be effective, data must be: (1) plentiful, as data-hungry models like deep learning-based models require large amounts of data for successful training; (2) high-quality, in the sense that the classification or labeling of the data to reflect the true underlying nature of the data must be accurate and the data must not be biased; (3) updated frequently with new data, since many natural processes are non-stationary and their behavior changes over time; and (4) pre-processed carefully, since real-world data is often dirty, missing values, and/or in various different formats.

Data for training AI models can be sourced from databases managed by organizations (structured and semi-structured data); APIs for web scraping or other specialized tasks; and data repositories. Low-code AI tools are emerging to further accelerate the B2B adoption process by reducing the time to build AI models customized to particular business applications, calling for collaborative human-centering design principles.

Providing access to the right set of labeled data is critical and challenging, since machine learning models are data-hungry and a significant amount of time and resources are spent on obtaining, preprocessing, and training prototypes of different model architectures. While some data may be collected easily for specific tasks, such as sentiment classifiers or name entity recognizers, probing for task-relevant labeled data in sufficient quantity and quality is not trivial. Applying AI tools directly to data-hungry tasks is a viable solution but will only work for simple and low-risk tasks like sentiment classification.

7.1. Data Quality and Availability

What we teach Cognitive AI largely comes from a narrow definition of big data. Therefore data quality and availability are the biggest blockers to widespread commercial use of Cognitive AI. Cognitive AI will be capable of learning from a wide variety of readily available unstructured content. However, the present foundational models can only learn from a small set of curated data, and they may only retrain with marginal improvements from a few specialized datasets — restricting its efficacy towards only "niche" businesses or areas of expertise. Advances in active learning, reinforcement learning, in-context learning, and few-shot or zero-shot training frameworks may ultimately take us there, but even then the resulting models will not match the abilities of a Specialist who would "naturally" possess the requisite expertise. When the Cost of "Real World" Re-Training is High, It Impacts Availability of Domain-Specific Models It is paramount to recognize that conventional re-training for most end-user businesses is not only very costly in terms of time, effort, and capital, but also fragile in its risk profile. In the present state of the art, for what may be seen as "plug and play" Cognitive AI capabilities, once they have been initial trained from large curated datasets, subsequent re-training (or additional fine-tuning) towards emergent specialized tasks or

use case is controlled via a purely technical process — a regression testing regime via rigorous, quantitative metrics of vast amounts of training data for task, with the associated success metrics for task fine-tuned along guidelines for potential task replacement that were also validated previously. The resulting huge amount of cost and effort required thus limits the number of nascent Domain Specialists having "plug-and-play" Cognitive AI capabilities solely available off the shelf for commercial utility (M. Stanley et al., 2024).

7.2. Data Privacy Concerns

Cognitive AI relies heavily on sensitive and private data. The data provides deep insights into people's characteristics and personalities, especially emotional and health profiles. It can come from diverse sources, both internet-based and provider-centric. The internet-based sources are social media and platforms where the users have a sharing culture. The provider-centric sources include integration of various user activity logs, transaction records on online marketplaces, user-generated data on collaboration tools, customer-provider interaction histories such as contact centers and live chats, wearable technology devices, mobile apps with location access, automated devices, etc. As the volume of data to train cognitive AI cores increases, the cost associated with private data collection and processing becomes significant. In addition to cost, access to user data is regulated under different privacy laws.

Violation of these legalities could lead to heavy penalties and may tarnish the reputation of the data-owning companies. With growing concerns about data privacy, users across the world have developed a fear of exposing their private lives and thoughts in the public domain. As a corrective measure, various technology companies are implementing features to overcome privacy concerns. For instance, some companies do not require all types of personal data to sign up. Both have built-in monitoring systems to let users know which app is using what type of data. Some let users delete their past data at any given time, while others let users stop location tracking and prevent eavesdropping by devices. Data privacy is one of the major concerns in advancing cognitive AI (J. M. Kizza, 2020).

7.3. Harnessing Big Data for AI

Big Data plays a crucial role in machine learning, which relies on the availability of a considerable amount of input-output labeled pairs to reach high accuracy. Cognitive AI systems often endeavor to model very complex relationships so that they can continuously learn from and adapt to changing situations. For instance, a system requires terabytes of specialized medical data to reach a level of accuracy such that it can safely make clinical treatment recommendations. A project used hundreds of terabytes of data to train their Recurrent Neural Network and achieved remarkable results in developing musical melodies. It is argued that unless a model is capable of dealing with an enormous amount of data, it will not be considered very crucial unless it can at least be a monthly online algorithm. Therefore, what distinguishes cognitive AI technologies from traditional AI highly relies on their demanding requests for both big data and humanlike

cognition. Furthermore, to be fully cognizant, cognitive AI should also be as efficient, cost-effective, and less time-consuming as human cognition at the same time.

The rise of AI is inspiring many companies and governments to invest heavily in research and engineering efforts, hoping to capitalize on this marketing trend. The success strived for by governments has not been as sensational; however, many have moved up to support basic developments and encourage innovation in many subdomains. Recognizing that machine learning – and in a broader sense – AI is the key to realizing humanlike cognition – it has been suggested that the establishment of a new infrastructure for AI, which utilizes big data and prevents model growth through cloud services, is capable of transforming how AI and machine learning are adopted and fostered. For instance, a novel machine learning company announced a significant investment with the ambition of providing mass services based on cutting-edge techniques (S. Manoharan et al., 2023)

8. Interdisciplinary Approaches to Cognitive AI

While Cognitive AI can be viewed as another novel endeavor led by computer and data scientists, psychologists, their colleagues from other fields of Cognitive Science, and philosophers have analyzed and discussed Cognitive AI's capabilities, merits and drawbacks, and weaknesses from their respective disciplinary perspectives. Their views are relevant for how Cognitive AI progresses. Exploring Cognitive AI from different angles can illuminate aspects of its operation, and the commonalities and differences with humans regarding learning and reasoning. Discussing multiple perspectives can also engender funding and broader collaborative models attempting to integrate diverse perspectives on Cognitive AI, leveraging the different strengths and weaknesses of each field.

A number of psychologists, especially cognitive psychologists and psycholinguists, have been analyzing the latest generation of LLMs in order to probe: (i) what the behavior of these models reveals about human cognition? And, conversely, (ii) what is the potential of these models to capture human cognition? This form of interdisciplinary research is called Modeling Cognition using AI. Several of these researchers have proposed a series of best practices that glue the field. These practices concern what particular AI model to use for which task, focusing on developing stimuli and tasks that minimize confounds at both behavioral and mechanistic levels, corroborating claims by linking behavior to internal structure and/or parameter settings of AI models, contrasting AI models' predictions against human behavior for specific experimental conditions that capitalize on the difference between humans and the model, and how the models should be understood when used in a cognitive modeling context. Notably, most accept the original stated mission of AI as defining our aim: creating programs that could, by virtue of their intelligent behavior, help us recognize and understand human intelligence.

8.1. Integrating Psychology and AI

Perhaps one of the oldest research programs in AI is its connection with cognitive psychology. The research established a close connection between machine intelligence and human abilities, the so-called physical symbol system hypothesis; many of the early models of human cognition were actually symbolic AI systems. It is only natural for both disciplines to bear on each other; humans are a sort of benchmark for any practical cognitive system, so a link between AI and psychology seems inevitable. However, as AI systems became more and more capable, these links began to fade; in particular, the ability of AI research to connect with what should be important experimental paradigms began to diminish, as there seemed to be less and less common ground to explore. What seemed to be an offer and demand equation had become a winner and loser equation.

In the past years, this trend has been somewhat corrected; it seems increasingly clear that, if we wish to explore human cognition correctly, we need proper partners in this endeavor. Psychology has developed advanced experimental setups and statistical tools that allow the exploration of phenomena from non-routine perspectives. However, what psychology lacks is a proper explanation and modeling framework for many of the interesting phenomena and events they highlight. Cognitive AIs are a proper debate partner because, having been trained for enormous amounts of data, they can indeed explore interesting facets of stimuli as varied as words, pictures, sounds, or even kinesthetic information (G. Miller, 2023).

8.2. Philosophical Perspectives on AI

The rapid development of AI systems can radically impact many aspects of everyday life. Builders of ever more sophisticated tools need to think carefully about their ultimate goals and potential outcomes. Ideally, designers will take into account beliefs about knowledge, agency, and social organization that sustained humanity for centuries. Philosophy is particularly concerned with the possible or actual relationships among sentients. There are many categories of sentience and many possible avenues of reasoning that engage their attention. Humans engage in a special category of reasoning called "philosophy," which, among other things, shapes the way we build our tools. More than any individual novel tool or technology, our capacity to synthesize and combine ideas about how we should relate to the world around us is what distinguishes humans from other sentients.

Patterns of philosophical thought track major turns in human history; these patterns are subject to empirical investigation and re-interpretation. Researchers in the history of philosophy seek to classify these turns, examine key works, authors, or schools, locate meaningful patterns of development, and identify profound continuities of structure and purpose. The founding principles of philosophy are not siloed, separate from knowledge and understanding in other fields. Rather, the reverse is true: all knowledge is somehow contained within philosophy, and particular branches of knowledge serve specific purposes by addressing particular classes of relationships or functions. Traditionally, these branches include ethics, aesthetics, epistemology, philosophy of science,

philosophy of language, logic, and metaphysics. Philosophy's sub-fields are not intended to be descriptively exhaustive (T. Metzinger, 2021).

8.3. Collaborative Research Models

Many stakeholders from diverse domains lie along the pathway toward more powerful Cognitive AI technologies. Building upon distinct foundational perspectives – from computer science, allied technology sectors, and the cognitive and behavioral sciences – these stakeholders span the range from commercial enterprise to national security, to societal services, and beyond. Thus far, there has been little appropriate dialogue across these domains about the opportunities, possibilities, challenges, risks, and ethical issues surrounding these future technologies. For example, the alignment problem has been most widely discussed in the context of human-centered Cognitive AI solutions for commercial use. Meanwhile, national security stakeholders often view alignment and control research principles more from the standpoint of adversarial modeling. However, the failure modes of these future systems can and should be discussed as a community of stakeholders. Serious research investment into Cognitive AI, particularly as it approaches the capabilities of a synthetic human-like intelligence, warrants clear and informative collaborative pathways among all vested stakeholder constituencies.

Rich interdisciplinary, pre-competitive collaborations of research leaders in associated fields may help accomplish this most effectively. AI researchers, especially in the language domain, have repeatedly been surprised by the emergent capabilities of Neural Networks as more and richer content has been introduced. Practical work at the intersection of cognitive systems modeling, framework-oriented research in Cognitive AI, and ethical modeling stands the best chance of success in advance of efforts addressing these challenges along existing institutional pathways (L. Floridi, 2023).

9. Future Workforce Dynamics in AI

Considerable political and social sanctions have arisen from the rapid introduction of generative AI technologies that can potentially replace jobs and replace humans with software. These technologies are expected to make certain types of job function obsolete, while generating new types of job functions that require unique cognitive services and capabilities. In the coming years, this dynamic shift will compel organizations to rethink their structures, including the number of employees needed compared to past expectations, how certain internal job functions that are being carried out today will have to be modified, and what new job functions will have to be developed and filled. The growth of the new types of job functions will be digital-native capabilities that will significantly evolve existing job functions focusing on data reasoning and intention thinking along with data categorization and enhancement. Rapid generative AI technology adoption has created increased demand for skilled talent that understand AI, prompting enhanced reskilling and upskilling initiatives. As demand for talent in the AI ecosystem continues to grow, educational organizations are leveraging partnerships with

businesses to provide AI-specific training programs and micro-credential courses to fill skill gaps across the workforce. Creating an inclusive workforce in AI and associated roles is vital for companies to unlock the technology's full potential and ensure equitable access to economic opportunities for all. Talent is vital not only for economic growth and productivity, but also for innovation. It is talent who invents and builds products that radically increase the speed and capacity of advanced technologies, and it is talent who apply these technologies to solve the world's biggest challenges. Powering the AI revolution will require a concerted effort to expand and diversify the talent pool by bridging AI skills gaps.

9.1. Jobs of the Future

As AI continues to advance, the demand for talent in associated fields is increasing, including AI safety, policy, research, and strategy roles. There is a need for transdisciplinary professionals who can conduct research on trusted, explainable, ethical, and privacy-preserving AI. These workers do not require a conventional background in computer science or statistics, although these skills are certainly important. Instead, they will probe how models work, why they work, what their limitations are, and when they fail. However, to serve in these roles, it is necessary to introduce accountability and oversight — maintainers are not enough. While large models can be quite resource intensive to produce, a much larger number of smaller, narrowed models may need resource-efficient accelerators, who ensure that these tools deliver trustworthy results to clients in specific domains.

Advances are also paved in natural language processing and understanding, computer vision, and other domains which allow us to build more accessible AI systems and tools for various sectors. This allows us to create more user-facing jobs, for instance, with tools that create personalized outputs for educators, teachers, or architects. Similarly, the convergence of AI and explorable crowdsourcing is potentially paving the way for new modes of collaboration between consumers and big tech AI corporations. Recent advances in human-in-the-loop models have allowed users to define, customize, provide feedback, and interact with the AI outputs. These tools have grown in popularity and allow users to extend the capabilities of the base large language models significantly. We may be seeing the emergence of a new role — the interactor. Fill-in-the-blank or zero-shot prompting have been the popular modes of interaction with these models until now (F. Pasquale, 2020).

9.2. Reskilling and Upskilling Initiatives

To successfully pave the way for these upcoming transformations in the AI ecosystem, large-scale reskilling and upskilling initiatives are essential. Governments, educational institutions, and large corporations are embarking on initiatives to enable people in this workforce to pivot and adapt to changes. Reskilling is the process of training an employee in a different field from the one that they were previously working in. Upskilling allows employees to better their current skill set. In AI, the growing

importance of reskilling and upskilling of workers and students will allow the labor force to remain relevant in this fast-changing world filled with technology.

The COVID-19 pandemic made apparent the major issues with education systems around the world. The expedited pivot to a remote-first approach to education has exacerbated some of the issues with traditional educational methodologies. The current education sector is not geared for a future where the majority of students will be learning online and not in physical classrooms. Various collaborative efforts by groups seek to close the gap and provide resources and knowledge for people to easily understand and learn about these technologies. Programs allow students to gain experience by collaborating with industry partners on industry-driven projects and addressing pressing problems around responsible AI. These initiatives and programs working with universities are aimed at producing a generation of workers who are equipped and ready for the jobs created by the rapid advancement of technology (H. E. McGowan et al., 2020)

9.3. Diversity in the AI Workforce

The need for diversity in the workforce is particularly true for the AI workforce, as these teams are creating the tools that will shape how we live in the upcoming decades and beyond. If our culture and our experiences are not reflected in diverse training data as well as in the teams collecting that data and creating the applications, there is a risk that this technology will be biased against us and will not be designed in ways to address our specific needs. This was initially highlighted by the application of facial recognition technology, which worked well for light-skinned people but had highly inaccurate results for dark-skinned people. When facial recognition cameras were being used to identify rioters at the annual Hong Kong protest, questions were raised about the accuracy and the need to reassess the use of that technology as a large number of protestors had black faces to hide their identities. A report concluded that many harmfully biased technologies driven by AI were deployed without deep understanding of their impact on already marginalized communities. These results were both exacerbated by a lack of diversity in the AI workforce and further cemented the need for diverse teams to not only assess and implement safe and just solutions in the development of AI but also to enable a shift in the way AI is created so that those fixed on solutions must also consider the consequences involved (R. Benjamin, 2019).

10. Conclusion

This book has aimed to explore the shifting ground upon which our relationship with AI rests. By taking a longer perspective on how AI models have been developed and deployed, their roots, their present-day role in shaping our media environment, and what this means for their future development, we hope to have opened up some new ways of thinking about the possibilities and limits of cognitive AI for enhancing social well-being. This is perhaps especially important today as there is considerable pressure from

multiple sides – from users, from businesses, and from researchers themselves – to accede to a cavalier rush towards a fully automated future. Cognitive AIs will be tailored to the full needs of human users and to the circumstances under which their relationships are enacted, only if human creators thoughtfully negotiate AI's role with one another over the short and long term. Here, then, are some of the most important lessons distilled from our exploration of this terrain:

1. Technology does not develop in a vacuum. No tool can be designed and deployed without prior understandings of what problems it is supposed to solve and who it is supposed to serve. While people building these tools today are applying a wide and surprising variety of developmental logics, only they know how closely these goals align with the tools' earliest users.

2. More generally, these tools cannot be designed without forethought on how they may harm those not responsible for getting them built, used, and maintained. These potential harms are many, from junking up discourse environments, to warping models of personhood, to enabling authoritarian control of citizens, to enabling harassment and bullying, to lock-out from various aspects of social, commercial, and communicative life.

References

Abbott, K. W., & Snidal, D. (2024). The governance of artificial intelligence: Legal, ethical, and societal perspectives. Edward Elgar Publishing.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity Press.

Floridi, L. (2023). The ethics of artificial intelligence. Oxford University Press.

Iansiti, M., & Lakhani, K. R. (2020). Competing in the age of AI: Strategy and leadership when algorithms and networks run the world. Harvard Business Review Press.

Kizza, J. M. (2020). Guide to computer network security (5th ed.). Springer.

Kurzweil, R. (2024). The singularity is nearer: When we merge with AI. Viking.

LeCun, Y. (2024). Towards artificial general intelligence: A path forward. MIT Press.

Manoharan, S., & Raj, P. (2023). Big data: Concepts, technology, and architecture. CRC Press.

McGowan, H. E., & Shipley, C. (2020). The adaptation advantage: Let go, learn fast, and thrive in the future of work. Wiley.

Metzinger, T. (Ed.). (2021). The Oxford handbook of philosophy of consciousness. Oxford University Press.

Miller, G. (2023). Artificial intelligence and the human mind. Routledge.

Min, A., Shin, J., & Lee, S. (2024). Cognitive computing in healthcare: Advances in AI for medical decision support. Wiley-IEEE Press.

Narayanan, A., & Kapoor, S. (2024). AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference. Princeton University Press.

O'Gieblyn, R. (2021). God, human, animal, machine: Technology, metaphor, and the search for meaning. Knopf.

Pasquale, F. (2020). New laws of robotics: Defending human expertise in the age of AI. Harvard University Press.

Ruder, S. (2024). Reinforcement learning in finance. MIT Press.

Stanley, M., Riccomini, R., & Stanley, J. (2024). Data quality in the age of AI. Packt Publishing.

Summerfield, C. (2025). These strange new minds: How AI learned to talk and what it means. Viking.

Susskind, D. (2020). A world without work: Technology, automation, and how we should respond. Metropolitan Books.

Susskind, R., & Susskind, D. (2022). The future of the professions: How technology will transform the work of human experts (2nd ed.). Oxford University Press.

Woolf, B. P., & Alemayehu, C. (Eds.). (2023). Handbook of AI in education. Springer.

Zou, J., & Narayanan, A. (2024). AI snake oil: Separating hype from reality in artificial intelligence. Princeton University Press.