The Nature of Information, Semantics, and Effectiveness for Artificial Intelligence and Cognition. Author: Kerry M Brown, Ph.D., 2021

Abstract

This manuscript puts forward claims to help address foundational gaps in understanding Cognition and Artificial General Intelligence (AGI), including the nature of Emergence, Semantics, and Information. This includes criteria for assessing true understanding in AI models. How symbolic reasoning conceptualizes phenomena is described. Without a subsymbolic perceptual level to generate concepts, there is no symbol grounding. Grounding requires dynamics outside of its own symbolization. Grounding forms the set of symbols used at the conceptual level. It is claimed that this role explains Semantics. This approach naturally leads to established research on Conceptual Spaces and has implications for Semantic Vector Spaces learned via Neural Embedding methods. It also has implications for Information Theories. A claim is made that Semantic Processes form Shannon-like microstates and macrostates, while Effective Processes constrain Semantic Processes. Unlike existing Semantic Information Theories, Semantic Processes are pre-informational. The claims provide perspective on the Mind. It is natural to conflate percepts with the modified version necessarily created when conceptualizing through explication. The 'Hard Problem of Consciousness' is related to this Percept/Concept distinction. Concepts are always subject to Eliminative Materialism. The nonconceptual properties of Percepts cannot be eliminated. Intrinsic and Extrinsic Emergence are distinguished. It is common to assume extrinsic emergent properties are intrinsic to the systems evoking them. This presents a challenge for proving intrinsic emergence in AI. However, criteria are proposed for claiming a theoretical system intrinsically processes information and grounds symbols. By leveraging the functional properties of Grounding, the criteria can be considered for actual systems.

Contents

The Nature of Information, Semantics, and Effectiveness for Artificial Intelligence and Cognition. Author: Kerry M Brown, Ph.D., 2021	1
Abstract	1
Introduction	5
Manuscript Style Notes	7
Definitions.....	7
Chapters Overview.....	10
Chapter 1: Emergence	12
Emergence as Heuristics	12
Discontinuity between Mechanistic and Emergent Concepts.....	15
Generating Heuristics.....	19
Localizing Emergence	20
AI-Automata Thought Experiment	22
Eliminative Materialism.....	31
Defining Percepts and Concepts	36
Complexity, Emergence, and Consciousness	45
Emergence and Mechanisms.....	47
From Intrinsic Emergence to Grounding	49
Emergence and Science	50
Chapter 2: Performative Systems	53
Potential Limits of Representation-Driven Learning.....	53
Emerging Veridicality.....	58
Amalgamated AI.....	59
Ethological AI.....	61
Conclusion	63
Chapter 3: Explicit Representational Modeling	66
Introduction.....	66
Moving away from Intrinsic Emergence	66
Explicit Representational Models (ERM) and the Description Gap.....	68
ERMs Lack Grounding	70
Why are ERMs Used?.....	74
Qualities, Attributes, and Concepts	82
Graphical Models.....	83

Conceptual Spaces	84
Neural Embedding / Semantic Vector Spaces	88
Perceptions and Conceptions in Conceptual Space Models	90
Mapping and Model Types	91
Chapter 4: Information, Semantics, and Effectiveness	93
Introduction	93
Shannon Information Introduction.....	93
Shannon Information Requires Predetermined Symbols	95
Shannon Information is Non-Semantic.....	96
Information is Emergent	98
An Alternative Approach.....	103
Semantics/Grounding and Semantic Information	105
Overview of Claims	105
Preface.....	106
Notes on Symbol Grounding	108
Semantic Information vs. Semantic Processes.....	111
Forming Microstates	115
The Role of Knowledge in Semantic Information vs. Semantic Processes	116
Symbols Cannot Emerge Semantics	118
Grounding as Model Forming: Prerequisites.....	121
Semantic-Level vs. Information-Level Interpretation	122
Information is Symbolic, Semantic Processes are Subsymbolic	124
Criteria for Subsymbolic and Intrinsic Processes	125
Some Considerations for the Criteria Proposed.....	136
Percepts/Concepts and Semantics/Information.....	138
Related Grounding Theories	141
Twenty Questions	142
Subsymbolic in the Limit.....	146
Measuring Grounding	149
Grounding as Understanding/Comprehension.....	160
Semantic Information.....	163
Semantic Compression.....	172
Agent Coordination.....	179
Some Alternatives for Characterizing Grounding or Semantic Information	181

Effective Information.....	182
Candidate Descriptions for Concepts Related to Information	192
Integrated Information Theory (IIT)	200
Summarizing the Relation between the Processes	205
Conclusions	208
Chapter 5: Conclusions and Future Research	213
Conclusions	213
Future Research	215
References	217

Introduction

I want to be clear upfront that I do not believe any current research program is on a clear path to Artificial General Intelligence (AGI). What I argue is that some methods capture critical modeling properties for Artificial Intelligence (AI) that must be incorporated into any new AGI approach. They will be key enablers for whatever is to come. As we inevitably reach a lull in AGI interest due to the limitations of current investments, understanding what legitimate insights were demonstrated will be important to ensure we do not turn towards paths that should already be recognized as dead ends. I will argue that these critical properties can be found, or at least approximated, within Artificial Neural Networks (ANNs), and their potential has been magnified through Deep Learning architectures. This should not be taken as a recommendation to keep moving down the path we are on. More computational power and data, modified loss functions, hyperparameters, etc. will not lead to AGI. New fundamental insights will be necessary. This manuscript aims to understand insights that some models already demonstrate, however dimly.

Deep Learning research continues to demonstrate increasingly impressive capabilities, at least relative to any current AI alternatives, including in Machine Learning. A common claim is that Deep Learning is fundamentally limited, and that we must incorporate past approaches that allow AI to reason with symbols. As Symbolic AI (colloquially known as Good Old-Fashioned AI or Classical AI) was largely interested in building AGI, many of the concerns revolve around the narrowness of what Deep Learning can learn.

Deep Learning's successes are typically chalked up to computational power and data access. While this is true, it masks the fact that Deep Learning is demonstrating some important aspects of intelligence. That is, this excuse fails to explain why Deep Learning's algorithms are as effective as they are with these resources. This manuscript does not champion Deep Learning. Instead, it focuses on this essence that Deep Learning has so far incompletely grasped and discusses its implications for AI and our own cognitive processes. These insights have been known in part for decades and built into the concept of Parallel Distributed Processing (McClelland and Rumelhart 1986). Current alternative paradigms do not resolve the major issues that Deep Learning has not addressed, and we are not therefore at a stage where achieving human-level intelligence is only a matter of effectively integrating existing methods. We must go deeper; we must commit to go back as far as needed to put us on a viable path. We must identify what is necessary to build AGI on solid foundations. To this end, the manuscript puts forward claims on subjects that most AI researchers may deem too philosophical to be bothered with, particularly in early chapters. The points made on these subjects are necessary to understand the claims made in the final chapter on Information, Semantics, and other areas more clearly related to building a foundation upon which AGI can be advanced.

Much of this manuscript elaborates on this essence by discussing what is not captured by Symbolic AI, which is extended to a category I refer to as **Explicit Representational Models (ERMs)** that includes probabilistic methods that model phenomena with explicit labels as symbols. Of course, the limitations of ERMs have been discussed at length for almost as long as

the field of AI itself. This manuscript continues in a tradition set by earlier researchers and philosophers questioning the assumptions in how AI is modeled. I will argue that we largely remain in a representational paradigm. Biological intelligence, including that possessed by humans, is not fundamentally representational. This is not to say cognition does not involve representation, but that representation is secondary to performing with the world.¹ One major issue with ERMs is their lack of grounding, which has been discussed at least as far back as (Harnad 1990). An objective of this manuscript is to provide further insight into how grounding can be understood, why it is important, and using these insights to relate a variety of other important concepts in AI and Cognition.

Of course, it is all too easy to criticize without offering possible solutions. A major aim for this manuscript is to elicit novel theories and models towards AGI by hopefully helping orient future research a few degrees closer to the right path. Example research questions and ideas are presented in the footnotes. These are meant to spur additional ideas and dialog.

To get there, we must go deep; committing to go as far back as needed to be on the right path, even if it means discussing Philosophy of Mind and novel algorithmic approaches. In the process, I believe the claims made are applicable beyond AI and hopefully provide a valuable perspective across the Sciences. Much of the manuscript provides thoughts on topics of interest beyond AI. AI could be thought of as a study of how to form something that can form models, which could help us understand how we also model our world.

The need this manuscript fills, then, is as an AI manuscript that confronts our assumptions in AGI research to help find new approaches to achieving it. In the process, interesting Philosophy of Mind claims are made. For AI researchers, then, I hope this serves as an opportunity to step back and to motivate the discovery and research of promising new methods. For philosophers, I believe thinking about fundamental AI can provide insights on our own minds and reality. Committing to understanding and addressing the fundamental issues in AI is equally useful for hard problems in Philosophy of Mind and Cognitive Science. Many of the questions we ask about AI are questions we have not answered about ourselves.

Investors and institutional decision makers should therefore take note too. Again, if this manuscript's claims are correct, then ideas that simply modify existing architectures or combine existing methods will not achieve AGI nor will new concepts that do not resolve the fundamental issues of the concepts they are intended to replace or augment. There are scientific and philosophical issues that must be addressed through theory and research. But there is reason for optimism if the claims and related research ideas help set a viable direction. After all, this may be the most important technology we ever build.

¹ According to (Lexico n.d.), Cognition is *"the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses"*. This leaves a lot of room for what it means to have thoughts, experiences, and sensations. In this manuscript, I take seriously the assumption that these functions must be intrinsic to a system for it to be deemed cognitive. I further add that to be intrinsic, the system cannot just be a model; it must be able to model, and therefore be a modeler. This is particularly important for understanding more specific requirements to achieve Artificial General Intelligence.

Most of all, I hope the claims in this manuscript lead to productive dialog and many fruitful collaborations. This manuscript had to hit on many different subjects, and I anticipate that this will leave opportunity for constructive feedback, including references missed and similar claims I should have been aware of. If so, then I hope to have integrated ideas of others in useful ways and provided new ways of thinking about existing theories.

Although it is discussed, this manuscript also does not attempt to solve the Hard Problem of Consciousness. If qualia serve some functional role, then, given its all-encompassing nature, it is likely to be fundamental to many of the problems described. My sense is that this manuscript discusses valuable concepts without resolving the nature of qualia. I do my best to separate qualia-dependent discussions from claims and research that do not require an answer to the phenomenal nature of experience. In the next section, I define terms used to help preserve this separation.

One final note: Some of the claims made may seem obvious to some. My experience, however, is that these claims are not fully appreciated, particularly in AI. If this manuscript serves as a reminder of established problems or claims, then I hope it will influence the dialog within the AI community and adjacent scientific fields.

Manuscript Style Notes

Blue bolded content highlights key claims.

Black bolded content is meant to make terms used easier to find when used again in later sections.

Footnotes in italics are example Research ideas or open questions that relate to claims in the main text. Some particularly valuable research concepts are also bolded in blue text.

Other points were also moved to footnotes to maintain flow. Reading them is recommended to improve understanding of the claims in the main text.

Many terms are capitalized to help distinguish them as core concepts in the manuscript.

Definitions

The following definition extracted from (Goertzel 2015) captures what the manuscript means by **Artificial General Intelligence (AGI)**: An artificial system (e.g., computer model, robot) that *“...displays the same rough sort of general intelligence as humans.”* This does not necessarily entail that the system be designed like a human or solving problems the way humans solve them. It also does not preclude the system from superseding human performance or solving problems humans cannot solve. What is important is that it can at least solve the range of problems that humans can generally solve. These typically fall into the category of “AI Complete” problems (Mallery 1988). These are problems we often envision AI solving, but they cannot because AGI has not been achieved. It is unclear how AGI can be achieved or whether any current AI programs are moving along a path that would in theory achieve it.

‘**Biological Intelligence**’ is used to refer to intelligence of living systems, including individual organisms. By default, the biological species in mind is the human. However, I did not want to discount other species when contrasting biological intelligence against AI models.

I will use ‘**observer**’ and ‘**receiver**’ interchangeably to refer to the agent that is receiving ‘**data**’ or ‘**messages**’ from the environment. Typically, ‘receiver’ and ‘messages’ are used in discussions focused on Information or Semantics, whereas ‘observer’ and ‘data’ are otherwise used. Note that the observer/receiver can refer to either a human who is observing an AI model, or it can refer to the AI itself observing its environment. ‘Environment’ here can refer to the outside world, including other agents, or to internal stimuli, such as an upstream neural network layer or the previous mental state of an agent.

‘**system of interest**’ is often used to refer to whatever is the object of study that is being referenced in the point being made. Often, the system of interest is an AI model.

‘**qualia**’ and ‘**experience**’ are used when referring to the phenomenal character of an experience itself. ‘Subjective experience’ has been used similarly (Tegmark 2017). Since I cannot imagine any experience that is not subjective in the phenomenal sense, I drop ‘subjective’, unless I felt a need to emphasize the subjective nature of the experience. Likewise, ‘intrinsic experience’ is used when it seems useful to emphasize the intrinsic nature of experiences.

(Harnad 1990), who proposed the Symbol Grounding Problem in (Harnad 1987), describes ‘**grounding**’ as involving “...*the semantic interpretation of a formal symbol system...made intrinsic to the system, rather than just parasitic on the meanings in our heads...*”. Grounding involves semantics associated with received symbols from the point of view of the agent/model of interest. In this manuscript, Grounding is defined as separable from the phenomenological properties associated with intrinsic experiences. This separation may not be faithful to the original meaning, however, it is consistent with how ‘grounding’ is sometimes applied today.^{2,3} Thus, Grounding is emphasized as a functional phenomenon (i.e., a means for coping with the environment), whereas intrinsic experience is used descriptively, with an emphasis on its phenomenological properties. The best way to think of grounding, as it is used in the manuscript, is: The *process* by which certain stimuli are used by an agent/model to evoke perceptions that can affect how an agent copes with aspects of the world. Grounding is still intimately tied to intrinsic experience. (Harnad 2007) claims it is a necessary but not necessarily sufficient condition for meaning. In this sense, it might be said that meaning requires both grounding and intrinsic experience, giving the latter a purely phenomenological role. Grounded processes must involve some different mediating processes between inputs and outputs that we would associate with perception, but it can in theory be measured while the intrinsic phenomenological properties of

² I do not justify this use *because* the term is sometimes used without claiming anything about an intrinsic experience. Instead, I believe it has been applied this way because grounding is itself grounded differently than just as the qualities of intrinsic experience. There is a semantic difference between the concept of grounding and the concept of qualia, and the former is commonly used without claims about the latter. Unlike ‘symbol grounding’, I would argue that ‘symbol qualia’ does not convey something perhaps like the activity of mapping a referent to a symbol. This is my justification for using grounding as the functional counterpart to intrinsic experience. I am unaware of a better term for the implicit properties that must accompany a symbol for it to be a symbol to the agent possessing those properties.

³ This biggest challenge for using ‘grounding’ here may be in addressing whether any digital model can be said to be grounded. I apply a different approach to this problem that I hope is worth consideration by those on both sides of the debate.

perception cannot. Ungrounded processes do not evoke perceptions given the same stimuli as their grounded analogs, and these perceptions can have a functional role in agent behaviors because of the role of grounding.⁴ Once all the functional aspects have been accounted for, all that is left are the inaccessible phenomenological properties. The latter have been isolated by what David Chalmers refers to as the ‘The Hard Problem of Consciousness’ (Chalmers 1995), and it is useful to maintain this separation for the purposes of this manuscript.

I use the term ‘**coping**’ as a relative and qualitative term for any dynamic between an agent and environment assessed with respect to some state or set of states. It is relative because the results of such interactions must be assessed against some implicit or explicit end state(s). For example, coping could be how much a given set of dynamics contributes to an agent achieving a desired grade on a math test, or it could be with respect to some implicit or ill-defined relationship to evolutionary history. The dynamics themselves do not constrain the end state(s) against which they can be assessed. However, given a chosen end state(s) to assess, the coping for that end state(s) is not arbitrarily measured. In practice, assessing coping is often difficult, particularly with partial observability or delayed effects with nonlinear, downstream interactions. Assessment requires assigning credit proportional to the causal role of the dynamics being evaluated. Coping is thus related to the behaviors evoked by context, be it external stimuli or a current thought or brain state. I will claim that achievement of a broad set of behaviors we would intuitively use as criteria for an agent to cope require not only Effective and Semantic Processes, but that these processes must be in certain forms. Typically, when I use the term ‘coping’ without referring to what ‘coping’ is measured against, I mean it generally. It is intentionally vague to avoid making narrow claims about objectives for AI agents. If helpful, when coping is used in the text, one can roughly imagine that it means how skillful or valuable most humans would evaluate behaviors are for the systems of interest.

Coping typically involves agents making use of what I will refer to as ‘**heuristics**’. These are biases towards certain ways of interacting with the environment, which can include ways of perceiving, reasoning, or acting. Heuristics could be learned or evolved. If adapted via evolution, a heuristic likely directly contributed to survival and reproduction. I will sometime use ‘Effective Heuristic’, which is redundant. This emphasizes the effectiveness of heuristics for agents over attempts to infer the “correct” answer or response for a given problem.

The qualitative concept of coping is in theory quantitatively measurable with **effective processes**, which, along with **semantic processes**, will be described in the ‘Information, Semantics, and Effectiveness’ chapter. In that chapter, grounding is claimed to be semantic processes, which is shaped by effective processes (In short: Semantics is Effective, such that

⁴ Note that ‘behaviors’ here is used broadly and includes internal processes. That is, Grounding is assumed to apply equally to dynamics associated with the external environment and internal stimuli. I see no reason to limit the concept to inputs and/or outputs from the external world.

Grounding enables Coping).⁵ Coping relates to grounding because the ability to cope often depends on how the agent implicitly maps symbols to external phenomena.

Emergence is commonly used in at least two distinct ways: A. In the sense of something being different than the sum of its parts (Anderson 1972); and B. something unfolding over time (i.e., a temporal dynamic process). Unless otherwise stated, this manuscript discusses type A. How new concepts emerge from lower-level mechanisms is both unexplained and foundational to cognitive processes and all experiences.

A critical distinction I make is between intrinsic and extrinsic emergence. **Intrinsic emergence** is the same as qualia or experience. It is what is felt by an agent from a first-person perspective. **Extrinsic emergence** can be understood from two reference frames. From the hypothetical reference frame of the system for which it refers, it is evoked by that system but is not experienced by it intrinsically. From the reference frame of the agent within which the emergence is evoked, it is experienced from the first-person perspective but as a property of something external to the agent (i.e., the system of interest). It is thus the part of an intrinsic experience that refers to phenomena perceived in the third-person. Much more will be said about intrinsic and extrinsic emergence in the manuscript. Note that the definition of Grounding used is claimed to be theoretically measurable, unlike any associated phenomenal properties, which are equivalent to intrinsic emergence.

Maps or Mappings, as well as **models**, are functions that relate one variable to another, while maintaining some relational properties between a set of variables. Something can be a map or model to someone without being grounded itself.

There are other terms used, some of which I will claim in the manuscript are equivalent to terms distinguished above. Since such equivalences are claims, rather than purely definition choices, I save these for the manuscript proper.

Chapters Overview

Chapter 1 proposes ways to understand Emergence. Much of the initial material is background and setting the stage for later more interesting claims in the chapter. While the claims in this chapter lean towards the philosophical, Emergence turns out to be very important in understanding and advancing AI. Chapter 2 suggests an increased emphasis on the performative nature of intelligence. Most of the claims are consistent with certain existing schools of thought in AI and Cognition. The specific claims in this chapter are important for situating intelligence in an appropriate context, which includes the importance of the concept of Effective Processes described in later chapters. Chapter 3 provides critical claims about the nature of explicit concepts in Science and AI. Again, the general points are consistent with claims that have existed for some time in AI but hopefully provide some new insights on problems with many types of AI models. To some degree, the earlier chapters are all preparation for Chapter 4, which provides

⁵ This definition for Grounding as equivalent to Semantic Processes means that Effective Processes directly influence but are not the same as Grounding. Justification for this separation and the dependence it creates between these processes are provided in the manuscript.

the most unique and specific claims put forward for a new understanding of Semantics, Grounding, and Intrinsic Information Processing. It is also the most difficult to explain and thus the longest. Due to the difficulty of describing the subject matter, certain claims are repeated to help understand their meaning through different contexts. Beyond the new claims made, this chapter also demonstrates how the claims made in earlier chapters might be applied to gain new insights in important scientific areas, including AI. Finally, Chapter 5 summarizes the main claims of this manuscript and includes discussion on future topics being developed that build on the ones that were ready for publication here.

Chapter 1: Emergence

Emergence as Heuristics

While there is no consensus on necessary and sufficient conditions for Emergence (Corning 2002), there is some agreement on certain properties. (Goldstein 1999) lists the following properties shared by emergent phenomena: 1. Features that are not previously observed in the complex system under observation; 2. Have coherence or correlation; 3. Observed at a global or macro level; 4. Dynamical (i.e., evolve over time); and 5. Are recognized by showing themselves.

I will generally distinguish between emergent levels and the underlying mechanism level from which the emergent levels arise. The features mentioned in the first property above will be understood as not being available at the mechanism level.⁶ The mechanisms evoke the emergent features but do not have those properties at their level. The easiest way to think about this is to conceptualize each level as states made up of some set of symbols that together describe the possible features at a given level.⁷ These symbol sets differ at different levels. For this reason, one level is not merely an aggregate of lower-level states, such as an average or probability distribution over them. When an emergent level is evoked in an observer, the change in the symbol set results in observing something “not previously observed”. The observer has a critical role in emergence. Even though the set of mechanisms that make up a complex system have a causal role in the emergent properties observed, the change in levels occurs in the mind of the observer. For observers, the emergent properties should be considered heuristics adapted to

⁶ The gap between these levels also has a role in how previously unobserved features can occur at the emergent level with respect to what has so far been observed at that emergent level.

⁷ I will later suggest that conceptualizing levels as states removes the very properties of interest of emergent phenomena. However, at this stage of the manuscript, it is easier ignore these issues to get other points across about Emergence.

cope with systems that we cannot fully sense and/or process at the mechanism level.^{8,9,10} We will see that these heuristics are more than just symbols.

With these in mind, everything that follows will assume that Emergence is a macro level generated given some mechanism level, and this macro level possesses properties that do not exist for the mechanism level. What follows are claims given this understanding of Emergence.

I take the view that emergence includes much more than the extraordinarily complex examples typically provided, such as life and economies. The experienced properties of water are emergent properties for the lower-level H₂O molecules. Indeed, it seems that all perceptions involve emergent properties. Therefore, **Emergence is important for Artificial General Intelligence (AGI) because it encompasses everything involved in biological intelligence as we know it.** AGI is the ability for a system without a biological brain (e.g., computer or computer model) to learn and solve anything a human can. This requires the ability to use the everyday concepts most humans know and apply. While AI may or may not experience emergence, it is pervasive in all human mental processes.

Effective mapping of emergent processes to environment processes involves Grounding. This grounding can be mediated by symbols to evoke emergent phenomena without being directly evoked by the environmental stimuli to which they map.¹¹ For example, I can evoke the grounding for a dog using the message ‘dog’ without having to see one in the environment.

⁸ My use of the term ‘heuristic’ here is associated with perceptions. Heuristics also cover many motor behaviors, both evolved and learned.

⁹ I do not directly address the other four characterizations from (Goldstein 1999) in the main text, but they are consistent with the claims made in this section:

2. The coherence/correlations in emergent processes also makes sense from a heuristic standpoint. If a system does not maintain or repeat itself, distinguished from the rest of the environment, then it is unlikely to be of any relevance. It could be said to lack persistent constraints. We attend to things that we cope with or can learn to cope with. This is important enough that humans can overfit stimuli, seeing associations where they do not exist, which may be a side effect of their being a higher cost of false negatives in some cases.
3. Effects of constraints that propagate through systems create global/macro patterns/features. Complex Systems typically contain structural constraints across mechanistic components that lead to phase transitions for the entire system. Since this involves a change in the properties of many components, it is often important for observers to distinguish these transitions. Different phases often require different coping processes (e.g., boiling water should not be ingested).
4. While most of the types of phenomena used to describe complex systems are dynamic, I assume that anything that is perceived as something other than its mechanistic components is emergent. This would then include everyday physical objects, such as rocks. They might have very different emergent properties but are still perceived as something with properties not exhibited at the underlying mechanism level.
5. The claim that emergent phenomena are recognized by showing themselves is somewhat vague but might be understood as the challenge in predicting or inferring the properties from knowledge of the components. Typically, there are no analytical solutions to derive how Complex Systems will evolve. However, complex systems are not entirely chaotic or random. The fact that they do evoke specific emergent features in observers suggests there are mechanism level constraints that we can leverage to cope with them.

¹⁰ I elaborate on the claim that emergence depends on an observer later in this manuscript.

¹¹ Assuming anything is fundamentally external to perceptual phenomena is inconsistent with the philosophical system of thought known as Idealism. This manuscript takes it as an axiom that there is something external. To reword everything to satisfy all possible metaphysical positions is impractical. However, even if Idealism were true, the experienced world includes the *perception* of external physical phenomena. I believe an Idealism-based science

Emergent processes are heuristic; in that they are selected (evolved or learned) to map to mechanisms in ways that tend to enable coping with the environment. Emergent phenomena can be thought of as models, and the adage, “all models are wrong, but some are useful” applies. How effectively they enable exploitation of some set of mechanisms ultimately determines how effective they are as heuristics.

To be effective, heuristics often must be reliably evoked by mechanisms that reliably lead to some subset of future mechanisms. This enables the explicit and implicit anticipatory behaviors that are hallmarks of intelligence (Rosen 2012). Relative stability enables intelligent systems to leverage constraints (e.g., correlations) for mechanisms through emergent variables (e.g., how a computer model predicts better than random that a sports team will win based on player statistics).

Emergent phenomena are often lossy with respect to the underlying mechanisms, reducing prediction accuracy relative to a theoretical system with complete access and the ability to process the underlying mechanisms. Most heuristics lose potential information and are irreversible (cannot be decoded back to the mechanisms that evoked them). Heuristics compress signals by reducing redundancy within and across signals. The tradeoff that determines optimal loss is between the capacity and energy required to process information vs. the effect to anticipatory performance given the loss.¹²

Although often sufficiently reliable to be effective, emergent concepts, as lossy heuristics, are not always fully reliable in how they map to lower-level mechanisms. Outcomes given applied heuristics will inevitably deviate from predictions with respect to the mechanisms as ground truth. Without full observation of the mechanisms, we can only be as precise as the heuristic variables that map back to them allow. Thus, emergent concepts are inevitably accompanied by some implicit uncertainty. Confidence is bounded when dealing with emergent concepts. **Lossiness and associated uncertainty, although difficult to fully specify, are important to keep in mind when modeling the processing of symbols meant to be grounded in emergent phenomena. Recognition of this all-encompassing emergence might help Artificial Intelligence / Machine Learning (AI/ML) practitioners to keep in mind limits in predictability associated with the concepts modeled and independent of the quality of algorithms or how much data is collected.**

Science is a process for determining maps/functions to reduce this information loss. However, as emergent concepts are often inherently lossy, this requires replacing them, often with new emergent concepts that are expected to be either less lossy, or to open paths that will lead to less lossy knowledge. Science thus develops new heuristics and new paths of knowledge, including the development of new scientific fields. Many scientists would claim it is impossible to fully reduce concepts to lossless (or minimally lossy) mechanisms, and I will make

would have to include this inherent perception of physical processes and the perceived discoveries that unfold from them. Thus, Idealists can simply throw the word “perceived” before anything that needs it, and keep in mind that these perceived processes seem quite fundamental to our perceived understanding of the perceived universe.

¹² Note that not all anticipatory behaviors can be described as explicit predictions. Many are measured with respect to behavioral adaptations, including inclinations towards behaviors, over a variety of time scales.

claims later for what specifically cannot be reduced to lossless mechanisms. More controversial would be the claim that the Sciences altogether aim at reducing emergent concepts to lossless mechanisms. For example, most Psychology researchers would not see their job as seeking atom-level explanations. My claim is that, as long as there is some loss between emergent processes and the mechanisms to which they map, there will be problems the Sciences have left to tackle. How we go about tackling these problems is what the Sciences do, and the associated uncertainty has driven what scientific fields have developed. This overall path happens to be extremely complex, and the nonlinearities hide a larger emergent driver from individual research and even entire fields.¹³ We will see that this drive towards less lossy concepts runs into several fundamental issues, including a loss of the very descriptions of interest.

Discontinuity between Mechanistic and Emergent Concepts

While lower-level mechanisms are responsible for emergent phenomena, they are not descriptively relevant when the explanation sought is at a higher level (which is almost always the case). We adapted the ability to generate concepts that can be predicted or inferred from other generated concepts to varying degrees of effectiveness. Consider the example of predicting what team will win a football game. One might argue that 100% prediction accuracy is possible in principle¹⁴ if we could account for all of the individual mechanisms and their interactions (including in the brains of the players) that could impact the game and follow their trajectories. However, while atoms may be fully responsible for what happens in a football game, determining who will "win" requires the emergent features because the molecular interactions have no way of explaining the emergent concept of "win". The football players, score, historical values, rules, etc., are all emergent phenomena.¹⁵ These emergent concepts are in the mind of the observer, which is the same language from which predictions and descriptions are being made. The language can be conceptualized as having a set of symbols that differs from those at the mechanistic level.

Language, including thinking, naturally depends on percepts, which I have described as emergent. Language therefore evolved with access only to emergent stimuli. Therefore, emergent properties are necessary for addressing questions already at an emergent level.

All higher levels are lossy emergent heuristic representations from mechanisms, but the mechanistic level cannot "explain" the higher levels. Lower levels may be best understood as causing emergent levels but not describing them. The causal association is not the same as the understanding of systems we build within levels, such as laws of physics, psychology, etc.

¹³ *Research: Consider if there is a way to quantify the relationship between scientific fields and loss, including what is gained when new fields are developed. Consider if Science's direction towards lower-level mechanisms mirrors entropy maximization, including through dissipative systems.*

¹⁴ Two arguments against Laplace's Demon (Laplace's_demon n.d.), or the ability in principle to predict with full accuracy, are the inherent randomness of underlying quantum processes and the need for infinite precision for predicting chaotic process due to extreme sensitivity to initial conditions (Prigogine and Stengers 1997). I set these aside here to make it easier to understand a claim about the nature of emergence.

¹⁵ Even prediction itself is always emergent, since there is nothing at the mechanistic level that can be *described as* prediction.

Modeling lower-level mechanisms can improve prediction at the higher levels, but this requires that the outputs can evoke the emergent concepts in the observer. In theory, this is not necessarily a problem. The point here is simply that the emergent level description is not satisfied by the lower-level description when what is being questioned/discussed is itself at an emergent level. It makes no sense to measure an emergent prediction *at the level* of the mechanisms.

I will claim later that a large subset AI research attempts to build explicit maps within and between emergent levels or assumes them in the models they build. This has been driven by misconceptions discussed throughout this manuscript. Identifying these maps is not only difficult due to partial observability of the mechanisms. These maps are themselves rarely describable because they are implicit.¹⁶ **This gap between levels creates an issue for mapping them: These maps exist causally, but there is no reason to believe that there exists some explicitly describable map waiting to be discovered.** I will refer to this issue as the **Description Gap**, where mechanisms cannot explain emergent concepts because the concepts we would use to describe them are lost with the concepts themselves at the level of mechanisms. This leads to a limitation not only for prediction but scientific understanding in general. Consider the following: If one maps the brain at increasingly lower levels of detail, at some level the concept of function will disappear. Function is an emergent concept. Even many evolutionary adaptations are heuristic concepts that emerge above the level of local biological circuits.

This Description Gap is related to the Explanatory Gap / Mind-Body Problem, or how neuronal dynamics give rise to the mind:¹⁷ If conceptual descriptions are emergent, including functional concepts, then this would prevent the ability to descriptively link mental experiences to the underlying physical processes.¹⁸

My general argument for the Description Gap is not that we as observers cannot build models that produce expected emergent effects. For example, an economic agent-based model could emerge expected increases in inequality or boom-bust cycles. I am also not claiming that one cannot find principles for how coarse-graining from variables to aggregates of variables can be mathematically explained (Pascual-García 2018) and (Hoel 2018) provide examples). What I am arguing is that the higher levels lose explanatory power when the lower-level dynamics are understood. For example, experiencing colors offer no explanatory value when we can measure

¹⁶ *Research: Determine a method/process that can learn more about the nature of implicit mappings. Are there invariant properties for them? Are there any properties that can be described beyond simply associating them with a statistical function given data (e.g., regression)?*

¹⁷ I would have preferred to simply reference this all as the 'Explanatory Gap' but the emphasis of the section, which is on the gap in descriptions across emergence levels, differs than an emphasis on how brain/body give rise to the mind. If this Description Gap concept can be accounted for by existing terms without losing what it is meant to emphasize, then it would be best to lose the term. The fewer redundant terms the better, so the same goes for any other terms I felt inclined to create in this manuscript.

¹⁸ *Aside: The question of freewill also seems to be an issue of emergent levels: Even if we did have a full account at the most basic mechanistic level of the system in question, including the entire history of what led to an event/action, we would be stuck at a description level that does not account for the heuristic concepts involved in decisions associated with freewill. Just as for the football game example, the level of explanation must be the same that the given agent employs.*

and use the wavelength information of the light that evokes them.¹⁹ Emergence is not just coarse-graining over the given system of interest. An observer's neural properties must be considered, and there is no reason to believe that anything besides associations could be found between these mechanisms and the emergent phenomena they underlie.

In the football game example above, a model implemented using lower-level variables could in theory lead to more accurate predictions of who would win a game but only in virtue of evoking the emergent concepts in an observer. The challenge is that the Description Gap makes it essentially impossible to identify how the lower-level processes must be modeled to evoke the emergent processes in any explicit manner. This is an issue for building AI to have emergent experiences similar to our own, since it means there are no explicit maps we can implement to model how emergent concepts we need them to have and to reason over can be evoked by the actual mechanisms that drive them.

There are existing ideas related to this Description Gap. For example, Sean Carroll (Carroll 2016) claims that different levels of abstraction/emergence each have their own limited domains of applicability. An example provided is that we perceive air as something other than a bunch of molecules. The way we perceive air works within the domain we typically use it but cannot be applied to infer that it could form a black hole if made sufficiently dense.

Black holes have not formed in the environments where we evolved. We have adapted a given level abstraction because it has been effective over our evolution. Therefore, emergence should always be considered in terms effectiveness (which I will describe more precisely in a later chapter). It also shows that it is lossy since many things could happen that would not be predicted due to their improbability of occurrence or effective impact on the agent. I claim that this is an issue for both Reductionism and Functionalism. The issue is not only that "more is different" (Anderson 1972): The explanatory information we seek in Science is itself at an emergent level, and therefore, the lower level cannot provide that level of information.^{20,21}

Just as atoms in a football game cannot explain factors we associate with winning because they do not occur at the same emergent descriptive level, **some evidence cannot be made explicit even though it partly informs on a fact.** If implicit knowledge cannot in principle be

¹⁹ This is not entirely correct. Color perception for humans, and presumably other animals, do not precisely map to fixed wavelengths, as discussed later. This is a good example of heuristics as it provides evolutionary advantages. However, the point holds if we were also able to account for the observer-dependent mechanisms, as mentioned in the text.

²⁰ *Research: What are the implications for Weak vs. Strong Emergence (Bedau 1997) if the explanations sought are themselves at a heuristic level? Note that the questions themselves must depend on lower-level processes. It is interesting to consider how questions depend on the lower level, such that questions about such dependences are always at a higher level than the dependence.*

²¹ *Research: Consider what an investigation would look like to better understand this Description Gap. Even though we have no direct way to assess how arbitrarily our emergent percepts mirror real world mechanisms, we can try to look for common dynamics that our emergence operations seem to exhibit with respect to commonalities in their underlying properties. Are there at least some approaches that would generalize well by making use of the regularities in mechanistic properties and human emergent processes? Maybe there are useful ways to perturb processes across levels? Can one identify and leverage some regularity where levels are more lossy as they increase levels, at least in terms of predictability, to infer the distance of the gap from other levels?*

transformed into explicit knowledge, then there could be evidence for something that cannot be transformed into explicit evidence.²² Note that variables at emergent levels are often nominal. They cannot be ordered. Yet at the mechanism level everything follows mathematical laws that can be mapped to a continuous phase space. Since there is no ordering for nominal variables, this is an example where there is no way to bridge this gap simply by applying mechanistic laws. There is no map from the bottom up for emergent variables that preserves the descriptions at both levels.

This all calls into question whether we can develop full theories of reality with explicit models. Here, an ‘explicit model’ refers to a model that preserves descriptions of phenomena for a given emergent level. However, while the phenomenal experiences associated with emergence are unexplainable from mechanisms, emergent phenomena can be **grounded** for mechanisms. The lossiness of the grounding determines how much emergent phenomena feels²³ explained by the mechanisms to which they map. In many cases, such as the football game example above, the gap may be vast between mechanisms, and the adapted explanatory level is less effective than desired. In all cases, however, this grounding is implicit. No explicit description satisfies the mapping from mechanisms to higher level concepts. **Implicit grounding is the only way to connect explanatory levels.**

A central claim, then, will be that this implicit grounded is how we cope with and understand the world. It cannot be excised without losing how non-mechanistic concepts are both evoked and what they evoke. One would be left with a structure with nothing to structure. Except at the lowest mechanistic level, it is unclear what it would even mean to understand the world without this implicit grounding evoked with symbols representing emergent phenomena. Here is the relation to AI: **If AI is to have any understanding, including at emergent levels, it will also require implicit grounding.**²⁴

The assumption that emergent phenomena have different levels of lossiness assumes that emergence is structured as a spectrum or hierarchical. If emergent concepts follow hierarchical principles, then one could in theory reduce the lossiness from emergence concepts to lower levels by mapping from higher to the closest lower emergence scale successively (e.g., psychology → biology → chemistry → physics). Indeed, this is the intuition underlying reductive science. However, a consequence of emergence-as-heuristics is that this assumed hierarchy may not exist, at least not as a “real” property holding between concepts. Like the heuristics themselves, it may itself be a heuristic to model them hierarchically, where relationships to the lowest level

²² *Research: Does this partially explain why grounded beliefs are sometimes seen as different from knowing? We may have what we take to be evidence in ourselves that we cannot communicate, even to ourselves, in principle.*

²³ By “feels”, I am informally referring to the implicit sense associated with the satisfaction that one has for understanding some phenomenon.

²⁴ **Research: Determine why in some cases lower-level understanding increases predictive power more than other cases.** For example, genomic data can provide more accurate diagnostics than basing diagnoses on symptoms for some conditions. Is this just due to being closer to the mechanisms or is it due to being less dependent on humans compared to, for example, trying to predict a football game (that is, football game outcomes depend on the actions of other humans who are not only highly complex adaptive but are leveraging their own emergent heuristics in games)?

is only defined by effectiveness, and the levels themselves are imagined. That is, not only are concepts heuristics, but they are layered and mapped across heuristic levels.

This is not to say that hierarchies should be dismissed; again, some models are useful. As a heuristic, it can be even contradictory and still useful. It is intuitive to model concepts as sets and subsets. There is a reason we use hierarchical concepts. It is entirely consistent with the claim of emergence as heuristics that hierarchical abstraction would form in some cases, based on a tradeoff in coping effectiveness and loss with respect to underlying mechanisms. However, it does not follow that there are necessarily logical principles that entail emergence is hierarchical (i.e., that emergence has “real” levels). There may be statistical regularities that in general hold as effective principles, many of which may be implicit themselves (e.g., all the green stuff over there is grouped into an object, and the other colors are background or different objects). The point is that the concept of a conceptual hierarchy, like other concepts, may be lossy, and is lost with any attempt to map them to mechanisms. As we will see, this creates problems for approaches that model the world with explicit concepts in formal, discrete structures.^{25,26}

Given the above, how can something like the use of emergent processes applied by biological intelligence be implemented in computational models? What if any AI model types might meet criteria for which we would conclude that they apply emergent processes? What are the criteria? This manuscript spends considerable time raising considerations and claims to address these questions.²⁷

Math and Logic often have no more ability to *explain* complex systems and concepts than understanding atoms in football players would help to predict who will win a game. We can in theory, and often in practice, precisely model emergent phenomena from simple lower-level mechanisms, leading to evocation of the same emergent phenomena in an observer. This is incredibly useful and is general practice in Complex Systems research. However, we must be careful to not conflate this with an explanation of emergence itself, particularly for building AGI agents, where we also expect them to possess, or at least act as if they possess, the emergent processes intrinsically. These processes demonstrate how to *cause* certain emergent phenomena to be evoked in observers but cannot *describe* the emergent phenomena at their own level. I elaborate on the implications for AGI later.

Generating Heuristics

Although it is not critical to the points above or below, it is worth mentioning that the pervasive role of emergent heuristics provides an opportunity to build AI that is effective beyond ways we can currently predict or possibly comprehend. As models, the space of possible heuristics may be essentially boundless. **Thus, we may be able to expand the processes that we use to develop heuristics through AI.** Statistical Learning, under the right conditions, might

²⁵ Research: Is there an effective process for determining the emergence level to model to observe higher-level emergent processes desired? Consider how to determine and model candidate selection principles.

²⁶ Research: Determine how to measure how close different emergent level concepts are to the mechanisms.

²⁷ Research: Finding ways to categorize heuristic types may be useful, even though the categories are themselves heuristic concepts.

already in some cases provide a means for discovering new heuristics. In a later chapter, I will describe criteria required for models to ground concepts. When these criteria are met, there is no obvious limit to what heuristics the models themselves might develop. In turn, these could be useful to humans, including to facilitate scientific discovery. Some heuristics could also be discoveries of existing but undiscovered ones used by biological agents, while others may even be imperceptible and completely foreign to our own evolved set of heuristics. All these possibilities are potentially useful.^{28,29} There are three types of heuristics that might be achieved by AI:

1. Grounding heuristics already known by human users (e.g., processes for walking)
2. Discovering new heuristics that can be understood and defined by human users (e.g., new but interpretable chess tactic).
3. Heuristics emerging implicitly in the model processes that are not understood by humans (e.g., unconscious mechanisms that cannot be explicitly mapped to emergent concepts).

Although some AI/ML methods are theoretically capable of producing all three heuristic types, they are in practice limited in all of them due to explicit modeling, predefined labels, and/or predefined objectives, as discussed in the next chapter. How might some heuristics differ from those currently generated for or by AI/ML? Clues are found in how humans cope with complex systems (e.g., how we construct stories given observations, for which current AI/ML collapses relevant complexity).³⁰

Localizing Emergence

We naturally assume emergent properties we project onto systems are owned by those systems, including for AI/ML models. All mental concepts are emergent and imagined. They do not exist "out there" any more than "justice" or "unicorns". It is obvious that colors do not exist "out there" nor does the feeling of temperature (e.g., cold), but they are imagined as if they do. There are often specific mechanism properties that evoke the emergent phenomena in observers but what is evoked is evoked in the observers. It is naturally difficult to remember that our emergent concepts do not exist externally from our minds. **There is an evolutionary benefit to perceiving them as if they are properties of the underlying external mechanisms with which we must effectively cope. However, it can also result in misattributing emergent properties internal to systems. This misattribution is particularly problematic in AI, where what an observer experiences must be distinguished from what the AI experiences.**³¹ This results in misunderstandings about emergence, where the system is assumed to intrinsically hold the

²⁸ *Research: Consider that density, temperature, and entropy are all emergent. Are there provable constraints on creating new, emergent forces and principles?*

²⁹ *Research: Determine how to optimally discover heuristics through computer models that could then be leveraged by humans.*

³⁰ *Research: Determine how a better understanding our own cognitive heuristics (many of which are unconscious) could help identify effective processes we have not yet applied in AI/ML.*

³¹ *Research: Consider what insights may be derived from neuroscience experiments that measure how neural dynamics change with different environment-controlled perceptual tasks.*

emergent properties, rather than stimulating their projection onto those systems by the observer. In some cases, we may even assign conscious properties to systems due to specific types of emergence projections they evoke. For computational models, we must take care not to project our own emergence concepts that the system may evoke in us onto the program outputs as if they are embedded within it. This issue has been discussed at least as far back as (Harnad 1990), which claims that the meaning for certain types of AI is parasitic on the meaning of the modeler. (Harnad 1990) calls this conflation a “hermeneutic hall of mirrors”.

To experience emergence (and therefore experience at all), the emergence must be intrinsic to the agent/system. For the rest of this manuscript, I distinguish between ‘Intrinsic Emergence’, which is experienced by the agent, and ‘Extrinsic Emergence’, which involves emergent properties about external systems but experienced outside those systems.³²

Extrinsic Emergence is a property of Intrinsic Emergence experiences. The emergent words you are reading now may be extrinsically projected onto the page, but they are only a part of your current intrinsic experience, not an experience of the letters or bits external to you. More generally, **Extrinsic Emergence is a phenomenal property of a subset of intrinsic emergence experiences that is perceived as a property of some part(s) of the environment.**

Again, extrinsic emergent properties certainly depend on the mechanistic properties of the external entity to which they are assigned, but these emergent properties should not be conflated as being experienced by the entity onto which they are projected. **Even if there was something inherent in external systems that could not be observed from the outside, there is no obvious reason to relate it to the extrinsic properties we project onto them. We have no evidence that the requirements we have defined for emergence (e.g., in Complex Systems research) bear any relation to requirements for intrinsic emergence. There is no known relation**

³² In searching to see if the term “intrinsic emergence” has been used, I came across (Crutchfield, 1994). There, intrinsic emergence seems to aim for the same concept for which I apply the term, so I use it here. There are some striking similarities in the overall claims. For example, the paper refers to “...*what machine class or language [an agent] implicitly is restricted to or explicitly chooses when making a model*”. This process of the agent changing machine classes is referred to as “innovation”. This process aligns with my description of Semantic Processes in Chapter 4, which I claim determine Information content received from messages and is required for grounding symbols in messages. I refer to this process as determining the indices for which to construct a Shannon-like model. Crutchfield’s approach is much more expressive than indices, since selecting/constructing machine classes involves changing structural properties. I think a consequence of aligning grounding to innovation would be to suggest that innovation is occurring almost all the time, at least for humans. This might seem excessive given an expectation that a selected model class should be semi-stable. Then again, biology includes many rapid and sustained dynamics, so rapid innovation is plausible. Moreover, we do not know what model classes humans apply, and the metamodel space may be continuous, where class change is a matter of degree with minimal energy costs. Also aligned with (Crutchfield, 1994) is the notion that any experimental data will have some randomness induced by the observer’s choice of representation. I associate this randomness with lossiness in relation to the mechanisms that emergent phenomena represent (this may simply be a rewording of the paper’s claim, but I am unsure how much the authors would like to generalize “experimental data”). Interestingly, this *representation* uncertainty could potentially be decoupled from the *represented* uncertainty, although I will not explore this. There are many descriptions in the paper where it is unclear how they relate to claims in this manuscript, including how much my use of term “intrinsic computation” aligns with the paper’s use of the term. I leave a full analysis to future study. Regardless, the article is full of insights. Like other key works referenced in this manuscript, they will likely prove critical for advancing a wide range of areas, including AGI.

between the Extrinsic Emergence an observer experiences due to a system and Intrinsic Emergence of that system.^{33,34} We will see in later chapters that this distinction matters for AI.

An example of extrinsic emergence is the gliders in Conway's Game of Life (Conway's Game of Life n.d.) (Figure 1). The emergence is part of an observer's experience when observing the gliders move during a Game of Life run and so is part of the observer's intrinsic (first-person) experience. This is intrinsically experienced as if the emergence belongs to the cellular automata with the property of being from a third-person perspective.

I must stress that Extrinsic Emergence is integral to our knowledge of the world. It is how we characterize processes and is particularly important in characterizing complex systems. It is valid to identify external systems with these emergent properties since the systems are (typically) necessary to cause the emergent properties in the mind of observers. Again, what is important here is to avoid assuming these system properties are the same as those that would be experienced by those systems, especially within AI.

Finally, note that nothing I have said so far proves that a given system does not have intrinsic experiences. Indeed, we have experiences, despite all conceptual knowledge *about* ourselves being extrinsic emergent. Only the perception itself is intrinsic to the observer. What I am claiming is that our emergent experience evoked by external systems is generally not evidence for what or even whether those systems experience emergence intrinsically. We will see in the Information chapter how we can mistakenly assume that the information we model in AI is what the AI modeled.³⁵

AI-Automata Thought Experiment

In this section, I attempt to clarify the difference between measuring intrinsic and extrinsic emergence with a thought experiment. Since Intrinsic Emergence is ubiquitous in our own cognitive processes, it may be important for any cognitive models, including AGI. However, it is unclear how we could provably generate it in AI models. If Intrinsic Emergence could be proved in other systems in principle,³⁶ AI may provide the best candidate to do so, given the level of control and access it provides.³⁷ Proving extrinsic emergence is easy since it is observed. This is

³³ Extrinsic emergent properties are often used heuristically to infer intrinsic emergent properties of others. We infer that other humans likely have intrinsic experiences based on how they can behave. However, this offers no obvious help for forming a scientific account to infer if AI models possess intrinsic emergent properties.

³⁴ *Research: While this would never provide proofs, if we assume that other humans have similar intrinsic experiences, then we could map behaviors of others to the same intrinsic experiences we experience given similar behaviors. The objective would be to build up a private understanding of the world by regularities found between certain types of extrinsic experiences and one's own intrinsic experiences.*

³⁵ As I discuss later, **this may help understand why we have issues explaining qualia. That is, when thinking about the concept of qualia, we convert it from its intrinsic form to something projected onto ourselves as an external system (i.e., a conceived self), where it is now no longer the intrinsic experience we are having, but instead an extrinsic property of our current intrinsic (i.e., quale) experience.**

³⁶ Note that proving intrinsic emergence within a digital computer model would have implications for the viability of a theory that the Universe is fundamentally digital (Digital_physics n.d.). Any insights for how AI could have intrinsic emergence, when we can fully describe their machine code, apply equally to a universe built with code.

³⁷ *Research: If we could provably create intrinsic emergence in a computational model, then, if the number of emergent property types can be increased over time, we could map the emergent heuristics to the fully known*

what we observe as different than the sum of the underlying mechanisms of the external system of interest. What is it we are looking for with intrinsic emergence? How would be it observed or inferred?

Since I cannot think of way intrinsic emergence for other agents could be observed directly, I will attempt to conceptualize behaviors that would seem to require a system to make use of information beyond what is available at the level of mechanism features. The goal here is to come up with as simple of a model as possible and see where that gets us. For this, I will use a computational system, where we know the underlying mechanism (i.e., bits), rather than with physical systems, which could be at the quantum level or something yet discovered. AI is also the main subject in this manuscript, and it is by default modeled computationally.

The approach is to require an agent without full access to the lowest level mechanisms of an environment to generate emergent concepts to achieve some desired objective. This requires designing a model of an environment sufficiently complex and interactive to allow an agent(s) to evolve something analogous to, say, basic vision.

For this, let us go back to Conway's Game of Life (Conway's Game of Life n.d.). This is a cellular automata model that changes states within a two-dimensional (2D) grid world. Each cell in the grid can be in one of two states (black = alive or white = dead). How the model evolves is completely determined by the initial conditions (layout of alive and dead cells in the initial game state) and the rules. There are four rules that determine what state a cell will be in by counting the number of alive or dead cells of its immediate neighbors. Different initial conditions will create different dynamics, including the one in Figure 1 below.

Figure 1. Game of Life. Left: Game state at time, t . Right: Game state a few time steps later. A series of gliders is moving diagonally down and to the right. Taken from: https://en.wikipedia.org/wiki/Conway%27s_Game_of_Life.

Despite the simple rule, Conway's Game of Life exhibit clear emergent properties. Just one example is what appears to be a series multi-celled pellets (called "gliders") that move across the grid. Given that the rules are local and consist of two discrete states, there is no preserved entity really moving. This is an example of *extrinsic* emergence. We as humans experience velocity for these pellets, but there is no reason to believe that the code or hardware that creates the model experiences anything of the sort.

The experience of these glider movements is an intrinsic emergent experience for us as observers. Since we can experience the changes as pattern-driven emergent concepts, such as velocity, we could in theory use velocity as a concept. Again, since the cells are driven by discrete

underlying reductive bit dynamics. This could be used to increase the number of grounded concepts for an AI model to advance AGI. Beyond AGI, such agents could evolve emergent pattern detection and maybe classification from extensions of cellular automata, like the design described in this section.

rules that turn them on or off, there is no inherent velocity in the game. Yet we can emerge the idea of velocity and communicate with each other about the concept. The thought experiment, then, is to replace the human observer with an AI model. While there are confound-related challenges in AI explicitly reporting what it observes, we can try to assess the perception of persistence and speed based on how much AI behaviors suggest the likelihood of anticipating future locations of gliders. The idea below then attempts to build a model that observes the Game of Life and achieves some objective for which we could conclude with some degree of confidence that it must have learned certain emergent features. More generally, the model must exploit higher-level invariant features and/or relations that are not programmed into the game itself. Here are some possible steps:

- Overlay an evolutionary or reinforcement learning agent-based algorithm onto the Game of Life Cellular Automata model. “Overlaying”, here, means the AI’s state-action environment mirrors the Game of Life grid. No interactions that affect the actual Game of Life environment are required.³⁸
- At each time step, the overlaid agent’s set of actions is to stay put or move to one of the four neighboring cells in the same grid as the Game of Life model.
- The overlaid agents must achieve some objective by evolving or learning the ability to account for glider velocity.³⁹

For ease of reference, I will refer to this thought experiment as “AI-Automata”.⁴⁰ What objective should we implement for AI-Automata? Let us say the objective is for an overlaid AI agent to move around on the grid to reach all four grid borders (i.e., the East, West, North, and South walls) as many times as possible⁴¹ within a finite time while avoiding ever landing in the same cell as a glider.

³⁸ Research: Consider if there are any advantages to instead instantiating agent interaction directly within the automata. This enables the AI agent to influence the automata themselves but could make the automata unstable in unpredictable ways while the AI is learning. A goal might be for the agents to change the automata patterns to other patterns by manipulating the local mechanisms, but without prior knowledge of how to do so. To do the test right, the agent must only have awareness of the emergent properties in determining its behavior, but with the effects occurring on the local mechanisms. To prove beyond some reasonable statistical significance level that emergence is utilized, the agents must adapt such that their performance maps better to the emergent properties than the local mechanisms across a variety of games. Due to the possibilities made available through interaction, this idea relates to the concept of Enactive AI, which emphasize the need for the AI to be interacting, and even shaping, its environment, in this case to increase robustness. Enactive AI and its relation to the other E’s (i.e., Embodied, Embedded, and Extended) is discussed in (Rowlands, 2010).

³⁹ Research: Consider whether cellular automata / agent-based models would exhibit more interesting extrinsic emergent properties if given real-world thermodynamic constraints.

⁴⁰ I apologize to readers here, since there is likely a much more concise and elegant way to make the points I am making in this section.

⁴¹ Specifically, all four boundaries must be reached in any order before reaching another border would count. For example, **East, West, North, South, North, West, South, East**, would all count (8 borders reached), while: **East, West, North, West, West, South, East, North** would only count 6 borders reached. This eliminates cheating by learning to oscillate between one to three easier to reach borders.

The agent-based model is designed such that they evolve/learn new sets of rules to guide their behaviors. Unlike the cellular automata cells, the agents can observe each state in full (i.e., all grid cell values) at any time step. This is so they can observe the gliders in order to make decisions at each timestep. For an evolutionary algorithm, this would include mixing and matching how different states determine agent actions given the rules and modifying the rules through mutation and/or crossover (i.e., splitting and recombining two or more agent's rules). Agents that perform better (reaching more boundaries in less time) have increased probability of surviving for the next round.

To help ensure glider velocity is considered, the agents must evolve/learn over multiple changes to the game. For example, the grid board could be rotated 90 degrees, creating four different direction scenarios. Other changes can also be made, such as varying the observation and/or movement rates of the agents. For example, in one scenario the agent can only observe the grid every two timesteps. This effectively doubles the velocity of the gliders. These help constrain the requirement for the agents to use the concept of velocity to make decisions. As a test, some scenarios or types of changes can be held-out to build confidence that the surviving/adapted agents use velocity information. Additional training can be made available by implementing other cellular automata with gliders, such as moving at different angles, having more than one glider path, etc.^{42,43}

There are two major questions of interest with this overlaid algorithm. The first is whether its design even allows for agents to evolve/learn something that convincingly makes use of velocity (i.e., can it learn a good strategy/policy?). If not, then the second question is, what else does the algorithm need? Does it need memory? Does it need to be able to explicitly abstract features? Does it need operators, such as adding, multiplying, averaging of features? We will deal with these questions in turn.

While I have not provided constraints on possible rules that can be generated to control the AI, what types of rules are allowed will clearly affect possible performance. For now, we will consider models limited to explicitly mapping actions to game states, where states represent the grid configuration (i.e., a matrix of 0s and 1s for the alive and dead cells) and its own location on top of the grid. Thus, there are no aggregation or averaging operators and it cannot explicitly condition on past time steps. We assume the model has the capacity to represent all possible states and evaluate actions to take with them, and it can be trained enough times to learn an optimal strategy (i.e., what action to take given a state).

Can such a rule set learn an optimal policy? Given a grid (i.e., not the set of rotated, etc. grids), every possible state is covered. Since the Game of Life rules are Markovian and deterministic,

⁴² This could simply be rules implemented in a grid to create customized gliders. Given the steps so far, there is no fundamental advantage to using the Game of Life as the environment. It is used here simply to contrast testing for intrinsic emergence against something with rich extrinsic emergent dynamics.

⁴³ *Research: What would be a fair statistical control to determine if the agents are likely using perceiving velocity? What about an agent that has access to velocity engineered in as a feature, or provably optimal rules with and without velocity as a feature, given no foreknowledge of the rotations that will occur? Is it possible to prove an upper bound on performance that is lower than if velocity could be used as an emergence-invariant feature?*

they will always exhibit the same transition from any given state. It would seem absurd to claim the algorithm requires intrinsic emergence, since the optimal policy could, in principle, simply be “stumbled upon” by chance. Incorporating a learning algorithm simply improves the chance of reaching it. There is no room for intrinsic emergent processes in this model. Note that the same is true for the model’s bit level representation. Ignoring our extrinsic emergent observations of it, there is no room for anything beyond the static discrete mechanisms at the bit level. There may be processes that the model evokes in us that are not explained at a given level, but such processes are extrinsic emergent processes. Even if we specify the model at an emergent level (e.g., alive/dead, white/black, up/down/left/right), those properties are extrinsic emergent. At the very least, we have failed to disentangle intrinsic emergent properties from the extrinsic emergent properties.

By rotating the grid or making other changes described above, we have implemented the possibility that a single state may have more than one possible transition. If so, then the optimal policy is more complicated. There is information that the agent cannot directly observe, which is whether that grid was rotated, etc. This creates a hidden variable that potentially limits the performance of a Markovian AI algorithm.

If the algorithm could directly leverage information about earlier timesteps (i.e., have memory), then it could do better. Memory of just the previous time step would likely help to distinguish grid rotations. However, time can also just be unfolded as additional features in the algorithm’s observed states. This exponentially increases the number of rules, but, again, in principle, the optimal solution can be stumbled upon.

A more interesting question is what is possible when the AI’s capacity is too limited to learn and/or store every possible state. Here, we could compare a model that represents only explicit rules that map actions to states to an AI that performs function approximation, as is done with Artificial Neural Networks (ANNs). However, it seems to be of no help to prove intrinsic emergence. If we again focus on the bit level, then, *at that level*, there are simply static discrete states that fully account for the strategy learned. For the same reasons as the model above, all emergent properties we would assign to it must come from us; evoked by the model, but not intrinsic to the model.^{44,45}

The upshot is that it does not matter how well the algorithm performs because we do not seem to be making any headway on leaving room for intrinsic emergence. This becomes particularly clear at the level of the bits making up the model states.⁴⁶ There are no features or processes that must be *intrinsically* emergent over space or time, since every model consists of static and

⁴⁴ I will argue later that, functionally speaking, this type of model is fundamentally different than the one above. The fact that we are focusing on a single level here will be important for claims about grounding and intrinsic information processing.

⁴⁵ Note that, while the black box nature of decision-making being distributed over neurons may limit our own interpretability, a separate algorithm could in principle be employed to explicitly map the ANN states to the game states without ever requiring representation of velocity as a feature.

⁴⁶ This issue directly relates to the well-known Chinese Room Argument (Searle 1980) and the Zombie Paradox (Chalmers 1996). AI would appear to use intrinsic information to make decisions when we know there is no room for additional information beyond the bits in the agent and environment code.

discrete features and actions without any additional requirement to point to anything beyond its own mechanistic rules. We can even incorporate averaged features or other functions, including explicitly computing velocity, but these just become other features that are sufficiently caused at the mechanistic level.

We also cannot point to indirect or implicit constraints learned by the models that affect their behaviors, because any such constraints are extrinsic emergent. A model may behave as if it has learned to wait two time steps before moving up when a glider is five cells away, but that it must experience something beyond the bits explicitly represented in the code requires evidence. Not only is there no evidence in the code but there is nothing in the code that requires such evidence. The behavior at the bit level is fully causally accounted for by the bit states. Anything else is extrinsic emergent.⁴⁷

I believe we can say that **any digital model is fully determined by the mechanistic rules at the bit level, even if some would appear to us as leveraging emergent concepts, such as velocity. As long as discrete mechanisms are causally sufficient to describe a model's dynamics, there is no need or room for it to possess intrinsic emergence at that level. Any other level is constructed by the external observer and thus is not intrinsic to the model. In fact, it seems as if emergence simply requires emergence. Once something is reduced, it cannot at that level create emergence elsewhere.**⁴⁸ Intrinsic emergence might still occur as epiphenomenal phenomena, but then it would also be unnecessary to implement in AI models.

There is simply no way to even comprehend what criteria could be met by a digital model to assume it intrinsically emerges anything. Not only is there nothing that comes to mind for how a model can manifest properties above and beyond its mechanisms (i.e., code), but it also seems certain that any attempt would add superfluous explanation at the level in which the model operates. Whether it is the Game of Life or the overlaid AI, anything in addition to the mechanistic rules and initial conditions would overdetermine the system evolution, in the sense that the mechanisms are causally sufficient.

I claim that **the problem of proving intrinsic emergence has to do with the Reference Frame of the observer, and that any phenomena that are intended toward from this reference frame become Concepts, which are fully determined by their lower-level mechanisms. In other words, emergence overdetermines any conceptual phenomena.** Note that the argument is not that the extrinsic emergent properties are any less real than other levels, such as underlying mechanisms. The argument is that these extrinsic emergent properties do not necessarily exist to the systems themselves (i.e., from their reference frame). This is the main conclusion I want to draw from this

⁴⁷ Also, since the algorithms involved can be incremented one step at a time, they can be paused in between for arbitrarily long periods in between increments without any effect to the results. While this is no proof, it certainly seems like something that would not evoke intrinsic emergence. Any purely computational example could be incremented in the same way. Of course, this is an extrinsic assessment. Perhaps from the reference of the algorithm, there is no pause. However, this not only assumes the algorithm has a reference frame, but that intrinsic emergence has some interesting independence properties with respect to real world physics.

⁴⁸ Of course, this creates a dilemma if us observers are fundamentally composed of bits or discrete mechanisms, then we too would be fully described without the need to assume intrinsic emergence by some observer external to us. As long as the observer has such properties, this dilemma continues *ad infinitum*.

experiment. Demonstrated behaviors might be inferred as intelligent and even grounded but tell us nothing about whether these observed models are capable of intrinsic emergence. The very nature of our particular means for observing and intending towards external objects, including models, eliminates the possibility of intrinsic emergence because this process would always identify emergent properties as overdetermining the observed system/object in question.

What is going on here? I believe the issue relates to something that will be a major claim in this manuscript, which are the fundamental properties of Percepts. I will claim that percepts do not inherently have intentionality. Instead, attributes of a subset of percepts have the property of intentionality and are what I will refer to as 'Concepts'. Concepts are about percepts but are also a property of (some) percepts. Unlike the foundational percepts themselves, concepts are fully eliminable to the mechanisms to which they intend. Here is the relation to the problem with the AI-Automata experiment: When thinking about or discussing these experiments we evoke the AI-Automata as concepts. We as external observers transform the Game of Life into a set of concepts, such as gliders, or even simply a model with complex dynamics. In so doing, we have re-presented the Game of Life in a paradigm that is fully eliminable to the underlying mechanisms (here, either the bits or the primitive Game of life rules for all cells). Thus, the issue in proving intrinsic emergence is inherent in how we experience and infer the world.

Importantly, this generalizes to any AI model (or anything external). In fact, it even covers the dilemma for how to prove intrinsic emergence in other biological agents, digital or not.⁴⁹ As we will see later, this means that intrinsic emergence cannot be explained from the third-person, including from mathematical or physics-based principles, because anything third-person is subject to Eliminative Materialism (Ramsey 2016).⁵⁰

Conceptualizations of phenomena refer to but do not consist of the emergent phenomena they conceptualize. That is, concepts always possess intentionality. Any conceptualization of an emergent property is a symbol of the emergent properties of interest, rather than the emergent properties themselves. The concept of something differs from the thing it conceptualizes. It seems that the actual phenomena of interest cannot be explained because explaining necessarily conceptualizes the phenomena, which, as symbols, cannot intrinsically possess what they evoke in observers (i.e., what we use them to symbolize). That is, analyzing concepts will not uncover the properties they reference. Through analysis, one only ever moves from one concept to the next, each pointing to something beyond itself that is evoked by but not constituted in them.

I see no way around this for Intrinsic Emergence. However, since we adapted emergent phenomena as a heuristic, I assume that it should be assessable in some way, even if we cannot

⁴⁹ For example, mechanical/causal processes, energy/quantum fields, etc. when conceptualized also can provide no explanation for intrinsic emergence. Intrinsic emergence overdetermines such systems just as much as it does for digital systems.

⁵⁰ *Research: If emergence is a physical property, then would energy dedicated to its novel properties be expended over and above some baseline for brain activity underneath it? If so, are there experiments that could be designed to control for and localize it (e.g., measuring energy loss that cannot otherwise be explained because we can calculate the amount of energy the system would otherwise have)? If identified, could it be used as criteria for deciding if intrinsic emergence is a property for other systems?*

make any claims about the possession of its phenomenological properties. The functional aspects of emergence are what I will associate with Grounding.⁵¹ I will put forward claims for how we might be able to define conditions for Grounding, and that Grounding is equivalent to Semantics. Fortunately, this seems necessary and possibly sufficient for building AGI. Although a full explanation for grounding must be reserved for a later chapter, we can briefly explore it for the AI-Automata thought experiment. What I will later claim is that Grounding involves determining (i.e., forming) models (i.e., macrostates) and their associated microstates given messages. This requires that the dynamics that form and ground symbols are themselves subsymbolic from the reference frame of the model. While we cannot observe a model implicitly assigning meaning to symbols, we can determine the conditions by which it could do so. This can be seen by how we as observers ground the received Game of Life dynamics. As observers, we form a micro and macrostate, including its symbol set (i.e., model indices), evoked by the game properties. This is associated with how we represent the game, and it underlies the emergence we experience of gliders, etc.⁵²

I will also argue later how we can make claims about grounding for different AI model types based on their properties. This suggests ways in which semantics must be modeled for AI, as well as intrinsic information processing. Thus, we could apply these criteria to the AI-Automata overlaid algorithm to potentially determine if it is grounding the Game of Life messages it receives, even if we cannot make any claims about intrinsic emergence.⁵³

Before leaving the Intrinsic Emergence dilemma raised by the AI-Automata thought experiment, I want to raise other possible explanations. One is that systems may need to be physical and/or continuous to experience intrinsic emergence. Perhaps the hardware running the code is the source of intrinsic emergence. However, since the transistor switches function as discrete processes, one would have to run experiments to isolate what there is beyond the discrete states that explains the agent's ability to generate intrinsic emergent concepts. This suggests that we can reduce the problem to whether continuous variables are required for emergence (although the system may also need to be physical, where continuous variables may be a necessary but not sufficient condition). The issues, however, seems to generalize beyond digital systems. (Carroll 2016) claims Panpsychism, where each given particle has some basic level of qualia or consciousness, must be false, because we know the degrees of freedom for every single type of particle. If each particle had, say, two possible conscious states (e.g., happy and

⁵¹ This would seem to end up making intrinsic emergence epiphenomenal to grounding, at least if the assumption is that digital systems are used and the claim that there is no room for intrinsic emergence in digital systems is preserved. This also assumes a non-Idealist reality. What would otherwise be considered epiphenomenal in a non-Idealist reality could potentially have no unique function and still be ontologically primitive.

⁵² I will say this repeatedly, but describing grounding as forming micro and macrostates is not quite what I have in mind. It may be best understood as a metaphor to ground in the reader what is more fundamentally an implicit process that I am forced to explain using explicit concepts.

⁵³ To be clear, my claim is not that the agents or cells in these models are necessarily grounded or that the agent-based model or automata itself grounds symbols in the same way that we as humans ground what we observe from them. However, the agent-based model or automata could potentially meet the criteria I will set forth to ground "messages" that flow through them.

sad), then that would add more degrees of freedom. We can transform this claim to cover a continuous distribution of states between happy and sad. I believe this calls into question how physicalism could account for additional emergent properties simply through interactions. It generalizes the argument above for conceptualized digital processes to conceptualized continuous processes. We will see that the Reference Frame of the observer must be taken into account.

Another argument might be that the emergence only occurs with sufficient complexity. When complexity is high, an external observer may simply not comprehend how the agent's learned policy maps to a behavior that would mimic intrinsic use of velocity. I believe this explains exactly why emergence evolved in observers but does not provide any actual suggestion for why it should create intrinsic emergence in the observed system. I must note here that Intrinsic Emergence seems equivalent to the Qualia from which consciousness is composed. With this in mind, the above generalizes to a claim against a common assumption about consciousness, which I equate to intrinsic emergence: **Complexity is an insufficient explanation for consciousness (i.e., intrinsic emergence)**. Degrees of freedom could be incrementally aggregated (e.g., sum, multiply) for two particles interacting, then three particles, etc., but there is never a point at which you get the degrees of freedom associated with emergent qualities beyond each particles' degrees of freedom. Unless there is a way to relate them to those preexisting degrees of freedom then consciousness or emergence cannot be invoked at any stage. Even for phase transitions, or any circumstance with complicated constraints that determine that aggregate's total degrees of freedom, the interactions would sufficiently account for them, leaving no room for additional degrees of freedom for intrinsic emergent properties.

The purpose of this was to contrast intrinsic from extrinsic emergence, particularly for AI. However, I must note one possibility not addressed above. The thought experiments above cannot disprove that the bit processes of the code affect some external conscious medium (like a consciousness "ether"). 0s and 1s could be perturbations of this all-encompassing medium, and those perturbations could give rise to intrinsic emergence. Indeed, it seems that there must be something more than just 0s and 1s. Given the claim above that Intrinsic Emergence is equivalent to the qualia, there is a simpler way to conceive the issue with digital systems through a sort of *reductio ad absurdum* argument: Imagine software, programmed in binary, that is conscious. It is conscious enough to be self-aware and discovers that it is completely composed of 0s and 1s. How could 0s and 1s ever create awareness of themselves? How could it ever fully explain the concepts by which the AI conceives it is made of 0s and 1s by using 0s and 1s? It would have to conclude that there is something not accounted for by the 0s and 1s, which would mean the 0s and 1s are not, at least by themselves, the fundamental components of its reality.⁵⁴ If intrinsic emergence is experienced for a system, then, either 0s and 1s cannot be the ultimate components of its reality, or there is something else along with them that is not yet explained.

⁵⁴ One assumption might be that we see how 0s and 1s can emerge higher level semantics and syntax by looking at the meaning and functions that the programmer programmed in it and that the compiler then compiled to 0s and 1s. However, this is extrinsic emergence. These meanings are ours; not intrinsic to the AI running from the compiled 0s and 1s. We project them onto the reality of which we sit outside.

This a more philosophical problem; that there must be something beyond mechanistic primitives may be consistent with Idealist theories proposing top-down universal consciousness (Kastrup 2019). However, this topic is beyond the scope of this manuscript. As mentioned at the beginning, I only want to go back as far as needed to address a path towards AGI. The purpose of the AI-Automata thought experiment was to clarify the difference between intrinsic and extrinsic emergence and to set aside certain philosophical considerations for the more tangible implications for AI that follow.

Eliminative Materialism

In the above sections, I alluded to the full eliminability of concepts to underlying physical mechanisms. The belief that everything in the universe can be eliminated in this way is known as ‘**Eliminative Materialism**’ (Ramsey 2016). Here, “eliminative materialism” means that materialism, equated to the fundamental reality of underlying mechanisms, eliminates the existence of emergent phenomena. Emergent phenomena are not explained by underlying processes but instead no longer have any explanatory power given the mechanisms. There is considerable debate on whether such a claim is defensible. In this section, I claim what I believe is the proper understanding of how Eliminative Materialism fits within reality. This will be important for claims made throughout this manuscript.

In alignment with Eliminative Materialism, **all extrinsic emergent properties disappear once a system is fully described by lower levels. This cannot be said for intrinsic emergence.** I have yet to find an example of an emergent system that contradicts this claim. **Once all emergent properties are explained away at lower levels, what remains is the intrinsic qualia in the mind of the observer. Since these qualia cannot be explained by the (now fully mechanistic) external systems, it forever remains out of reach of materialism.**⁵⁵

To say extrinsic emergence is eliminable is to say that the emergent properties that reference external systems eliminate in principle to underlying mechanisms. Thus, the gliders, as properties about emergent phenomena evoked by Conway’s Game of Life, can be eliminated to mechanisms in the program code. If that full set of mechanisms were fully conceptualized, they would be casually sufficient; eliminating any of the emergent phenomena we might conceptualize about the game. So too, Psychology concepts can be in principle eliminated to Biology and ultimately

⁵⁵ Some Philosophers of Mind claim that qualia themselves are illusions. However, it seems to me that, at least for the ones I have read, they confuse the idea that the concepts formed from qualia are illusions with the idea that qualia themselves are illusions. Claiming that what we perceive are illusions is entirely different from claiming that experience itself is an illusion. The former is a claim for beliefs about what is perceived, which is a conceptual notion; it is *about* what is perceived. The perceptions experienced just “are”. This conflation is demonstrated in the claim that there being a self is illusory being an example of saying qualia are illusory (Blackmore 2005). The self being illusory does not entail that the experience we would associate with self is an illusion. Qualia is equivalent to experience, so if having experiences is not illusory, then neither are qualia (thus, I believe the claim, “I think, therefore I am”, actually assumes too much in assuming there is a real “I” that can be claimed to have the experience itself, rather than being a property of the experience.). Of course, as discussed in this chapter, the fact that the act of philosophizing converts percepts into concepts makes this conflation even more difficult to detect in one’s reasoning.

to Physics concepts.⁵⁶ However, the intrinsic emergent phenomena are not eliminable. The difference is that extrinsic emergent phenomena, although fundamentally properties of nonconceptual intrinsic emergent phenomena, are conceptualized properties. It is this conceptualization that makes them eliminable without addressing the more fundamental intrinsic properties associated with them.

If atoms are colorless and odorless, then how does color and odor exist, even if only in the mind? Neurons themselves are also made up of colorless and odorless atoms. There is no clear path to emergent properties from mechanisms, even if emergent properties are assumed to be mental abstractions constructed from mechanisms. This, of course, is the well-known Explanatory Gap (Levine 1983). Supervenience (Davidson 1970) proposes that higher-level phenomena supervene on lower-level ones if any changes in higher-level phenomena entail changes in the lower-level ones. While this provides a relationship, it does not explain how higher-level concepts arise from a fundamentally atomistic universe but only what drives changes in those concepts. That changes in qualia are determined by changes in neural states (or down to even lower-level mechanisms) is not surprising. What is unexplained is how qualia could ever arise from the physical mechanisms. Eliminative Materialism thus suggests that the emergent properties are illusory. There is no relationship to explain once we understand the mechanisms. My claim is that this is true only for extrinsic emergent properties: **Eliminative Materialism is indisputable but only in the extrinsic emergent domain. What is eliminated are the extrinsic emergent properties of an intrinsic experience.**⁵⁷ **The fundamental intrinsic experience, which is nonconceptual, is not eliminable.** The issue with the debates on this subject is due to the assumption that concepts put forward by both sides are the same as what they represent.

⁵⁶ The assumption I make here is that eliminative events do not only occur at the lowest level mechanisms. Lower-level scientific domains eliminating higher-level ones is difficult since higher levels often have less complete theories. Thus, “in principle” may be too strong, since there is no reason to believe emergent heuristic concepts, including non-mechanism sciences and their objects of study, are completable. Furthermore, Psychology has a heavy dependence on the environmental, including abiotic phenomena for which their underlying Biology-level causes would need to account. This does not forbid the possibility that concepts can be eliminated by lower-level concepts within emergent partial domains.

⁵⁷ I elaborate on this claim in a later section in this chapter.

Concepts of phenomenal properties are eliminable but the phenomena, when understood to be fundamentally nonconceptual, just “are”.^{58,59}

Note that Extrinsic Emergence **is all that Science addresses when conceptualizing our world. The consequence is that, as more is known about extrinsic emergent phenomena, it gets eliminated down towards underlying mechanisms, but the intrinsic emergent properties are not explained. They are irreducible, not because they are illusory but because they cannot be conceptualized and remain what they are intrinsically.**

It might seem to some that nothing intrinsic is left to experience once the extrinsic emergence is eliminated, but this would not explain why there ever any intrinsic emergent experience in the first place.⁶⁰ Without the distinction between intrinsic and extrinsic emergence, Eliminative Materialism has a problem in that there must be something to be eliminated. Even if one claims it is imagined, there must be something that gets unimagined. There is something experienced that differs from its mechanisms. Otherwise, we would not know the very difference between the experience and its reduced elements that we are trying to explain. To eliminate the

⁵⁸ There may be some confusion about why extrinsic emergence is eliminable and not simply concepts, particularly because the emergent qualities associated with any experienced concepts would seem to be the intrinsic nonconceptual properties I am arguing are not eliminated. The larger claim is needed to address specifically what one asserts belongs to a given system, and thus what Eliminative Materialism claims is eliminated. Consider Conway’s Game of Life, where the mechanisms are the bits of the game dynamics (or the bits for the rules and initial game state). Intrinsic Emergent properties are not accounted for in this set of mechanisms but only in the trivial sense that the observer’s mechanisms are not included in the set of mechanisms considered. Even if we do not consider the observer’s mechanisms, it still holds that the extrinsic emergent properties that one naturally *assigns to the game* are eliminated, because what is at question with eliminative materialism is whether those phenomena are real properties *of the system itself*. This is the sense in which I am using the term ‘eliminated’. It is not a difference about the nonconceptual content involved but about the assertion eliminative materialism is making (i.e., the reference frame). Our intrinsic experience as observers, in contrast, is not something that is assigned as a property to the game or any other external system in question. Also note that, even if we did scope the mechanisms to include the observer’s neural dynamics, this converts those dynamics to extrinsic emergent phenomena, thus also becoming eliminable. This would not address the more fundamental intrinsic emergence underlying those extrinsic emergent properties. Now, the reason I do not just claim that Concepts are eliminable is because, while concepts are necessarily extrinsic emergent, and thus eliminable, extrinsic emergence is not fully conceptual in the sense that the eliminative assertion makes, which is about the full extrinsic emergent property. If I have a nonconceptual experience of gliders, then when I claim that the game’s mechanisms eliminate them, I am making a claim for that full experience. This is a technical difference, since such a claim is necessarily conceptualized. It is only needed to address what is being claimed in the assertion about what belongs to the system in question. Specifically, the same nonconceptual phenomena that underlie a concept is both eliminable as extrinsic emergent properties and non-eliminable as intrinsic emergent properties. This technicality only matters for claims about Eliminative Materialism. For all other claims in the manuscript, it should be fine to simply assume Concepts are what are eliminated, and that they are always extrinsic emergent. It is really the Conceptualization that makes the full extrinsic emergent phenomena eliminable in the specific sense discussed. This is why the claim is for Extrinsic Emergence, but the justification for the claim is all about Conceptualization.

⁵⁹ I said above that the Explanatory Gap relates to how mental (emergent) properties arise from physical processes. The problem seems to be in the question “how”. To answer “how” is to conceptualize what is being explained, which converts it into something else that is merely about it; a percept into a concept referring to it, which is always eliminable.

⁶⁰ This is so obvious (seemingly self-evident) that the fact that some make this claim makes me wonder whether I am missing something profound for why intrinsic emergent properties would also disappear.

experience from some chosen set of the universe, only pushes it to some other plane arbitrarily determined to be outside of reality. There is simply nowhere to eliminate intrinsic emergence (qualia) to. A claim that qualia is illusory only applies to the *concept* of qualia or the *concept* of a given quale. The concept of qualia is readily demonstrated to be illusory in virtue of not being the same as what it conceptualizes. This does not make what it conceptualizes illusory. It is this, the intrinsic emergence, that critics of Eliminative Materialism rightly cannot fathom being eliminated.

Of course, suggesting eliminative materialism and intrinsic emergence are both valid feels unsatisfactory, as we would like to have a single theory to describe reality. This is the only way I can see to avoid the issues described above. If a single theory were necessary, then this would seem to favor the intrinsic emergence as fundamental, as it could cover the eliminative processes of external systems as being a mental process of the observer. That is, not only can I eliminate everything through reduction but my own experience, but my ability to reduce anything else is fully embedded within my *experience* of this system I have reduced.⁶¹ Eliminative Materialism seems obviously true for any case I can think of, and yet, it cannot explain the fact that I am able to have an experience (qualia) with emergent properties, without which, there should be nothing to eliminate in the first place. Even the concept of 'Eliminative Materialism' is emergent, and so must eliminate itself. However, to explore this further would take focus away from the implications for extrinsic vs. intrinsic emergence that follow, particularly for AI. The purpose here is to properly understand intrinsic vs. extrinsic emergence for addressing AI requirements in the rest of the manuscript. To this end, I make a few more relevant claims before ending this section.

Extrinsic Emergent properties can only be eliminated in principle. At our own emergent level, we cannot fully map them to lower-level mechanisms. This is simply a restatement of the Description Gap discussed earlier. It is important to keep in mind that we not only evolved extrinsic emergent heuristics, but that they dominate our mental life, including as we conceptualize requirements for future AI.

As I have already stated, Extrinsic Emergence occurs through Intrinsic Emergence where the intrinsic mental concept is about some perceived external emergent phenomena. Just as the subjective realm contains the objective as a particular type of subjective experience, and as discussed more in the next section, how concepts are a type of percept (a percept about a percept), so too intrinsic emergence contains extrinsic emergence (e.g., intrinsic projecting of

⁶¹ Note that something like experiencing a concept about my own experiences (e.g., considering the idea that I am happy) falls into the category of extrinsic emergence because, even though it is attributed to myself, it is perceived as external to my current perceptual experience. Intrinsic Emergence is a more foundational concept about the frame within which experience occurs, not a boundary between the types of subjects to which experiences refer.

novel emergent properties, such as an observer perceiving a couch instead of the underlying mechanisms). We will see how this all relates to first- vs. third-person perspectives.^{62,63}

If extrinsic emergence is a special case of intrinsic emergence (i.e., a perceived property projected onto an extended object that is intrinsically conceived), then this may be why it is difficult to explain if/how something "emerges" outside the mind of the observer who perceives it; like all perceptions, it is necessarily intrinsically experienced. Again, if extrinsic emergence is a property of intrinsic emergence, then intrinsic emergence is more fundamental, further suggesting that eliminative materialism is a property of the more fundamental phenomenological experience (i.e., qualia). The fact that any quale (that is, intrinsic experience) itself is eliminable when conceptualized (i.e., discussed or thought about as a quale) may be why eliminative materialists believe it can be considered an illusion. The trouble arises when we place concepts as fundamental over percepts. Concepts must be properties of percepts. The only claim I can imagine for making concepts fundamental is that concepts are required for shared knowledge: There is no science or philosophy to discuss without concepts. It may be true that we need concepts to structure our understanding of the world, but it is not a given that this structure is fundamental, rather than a means to constrain the underlying percepts evoked by it. I will argue that the relation between symbols and grounding is that the symbols constrain the grounding. This way of describing the process suggests that the structure, which we could call our conceptual model of some domain (i.e., a universe), is empty without the grounding. If so, then concepts are not fundamental but instead just necessary for how we share information (including internally as explicit thinking).

This has implications for the nature of Semantics, which are associated with Intrinsic Experiences. Since Concepts are fully eliminable, if concepts could be grounded using only other concepts, then the semantics evoked with them would be conceptual and thus eliminable. However, there is no Intrinsic Emergence (i.e., experience) without associated Semantics (i.e., grounding). Since Intrinsic Emergence is non-eliminable, Semantics must also be non-eliminable, and thus cannot arise from only other concepts.

If the real world out there is conceptualized in our processes of deliberately attempting to know it, then the fact that such knowledge must be grounded non-conceptually to mean anything to us (i.e., semantics) can be seen as a reason for the Description Gap. In many ways, this Description Gap is equivalent to the Explanatory Gap. The latter is typically discussed in terms of the gap between mind and matter. The Description Gap extends the problem to a gap in description between the concepts we use in thinking, language, science, etc. and underlying

⁶² Admittedly, these all point towards a common theme, where consciousness unfolds physical properties. It may suggest a way of explaining the world starting with intrinsic emergence and then how it evokes the intrinsically emergent properties of phenomena and laws perceived as "out there". Such a claim is not necessary for the AI-related claims in this manuscript (or at least is beyond the bounds of what this manuscript addresses), but it would be disingenuous not to mention it.

⁶³ Given that extrinsic emergence is a special case of intrinsic emergence, "intrinsic" emergence could simply have been referred to as "emergence" without the qualifier. However, later discussions about the nature of emergence seem easier to follow with the added emphasis.

mechanisms. **This moves the Explanatory Gap away from a focus on physical vs. possibly epiphenomenal phenomena, to one that directly affects our knowledge, including the very models we are building to understand our world and the entire field of AI. Unlike the Explanatory Gap, which focuses on consciousness in general, the Description Gap can be convincingly claimed to matter for how we make and assess models and drive learning in AI.**

In summary, conceptualization converts emergent experience into extrinsic emergent phenomena (from the reference frame within which they are concepts) and therefore makes them eliminable (in the eliminative materialism sense). The fact that extrinsic emergent properties are eliminable has consequences for other areas. For example, when we conceptualize ourselves or other agents, goals disappear once we know the full causal mechanisms. Even notions such as Freewill and Purpose are imagined if we can in theory understand all the underlying true efficient causal properties of a system.⁶⁴ This can also be related to the phenomenon whereby the intelligence of AI we perceive seems to disappear once one understands the underlying mechanisms.⁶⁵ In all of these cases, the distinction between intrinsic and extrinsic emergence is critical to keep in mind. Eliminative Materialism applies to examples of *extrinsic* emergence.

Defining Percepts and Concepts

An important claim, alluded to earlier in this chapter and related to the Description Gap, is the distinction between Percepts and Concepts. The distinction was used heavily in the section above. This section focuses and elaborates on this difference. **Concepts are Percepts about Percepts, or rather, Percepts with properties about phenomena. Percepts are fundamental, but Concepts are ubiquitous in all thinking and communication processes because there is no way to think about or discuss Percepts without converting them to Concepts.** Every mental experience is perceptual,⁶⁶ but some perceptions are *about* perceptual phenomena⁶⁷ (i.e., concepts). Whenever one is thinking or talking *about* phenomena, they are in the realm of concepts.

Percepts could be said to be equivalent to qualia (i.e., intrinsic emergence or experience).⁶⁸ A percept is a quale, and a quale is a percept. **Concepts always include ‘aboutness’ and have truth values associated with the phenomena they are about. Thoughts about intrinsic**

⁶⁴ Metaphysics, including for Morals and Ethics, are eliminated as well, leading to striking similarities with claims in (Wittgenstein 1922). However, this leads to logical dilemmas with what I believe are incredibly interesting implications. I had more to say on Value but ultimately decided it is beyond the scope of this manuscript.

⁶⁵ This suggests that at least some of us assume a level of emergent decision making in our criteria for Intelligence.

⁶⁶ By “percept”, I am referring to any experience, not just those evoked directly by the external world.

⁶⁷ “perceptual phenomena” is redundant in how I use both terms. All phenomena occur in virtue of being perceived and all perceptions are phenomenal. I felt there was room for misinterpretation by using either alone here.

⁶⁸ C.I. Lewis (Lewis 1929) distinguishes what he coined as ‘qualia’ from concepts. This aligns with what I am calling percepts. One distinction I make is that conceptualizations are also perceived, and thus properties of percepts. Lewis also claims that qualia do not in themselves constitute knowledge. While this seems correct at their own level, I believe they do *conceptually* count as implicit knowledge, since they are influenced by experience (e.g., changes to a neural network underlying them will also change their properties). This understanding will be clarified later in this manuscript.

emergent properties (i.e., making it the object of consideration) turns them into concepts about those percept properties (e.g., perceptual qualities of a dog converted into the concept of a dog). However, Concepts are still part of Percepts; Percepts can contain Conceptual properties. **When considering Percepts with Conceptual Properties, one can take the perspective of their Conceptual or Perceptual nature. These perspectives can be thought of as two different levels or Reference Frames.**^{69,70}

Percepts do not fundamentally have "aboutness" (i.e., intentionality).^{71,72,73,74,75} They can have aboutness as an attribute only for any concepts they may contain. This may be a controversial claim, given that a hallmark of conscious experience is considered by some to be that it has intentionality.⁷⁶ I am claiming this is not a necessary condition and is not fundamental

⁶⁹ I believe that describing these as different reference frames is more appropriate. However, using this term becomes awkward in some cases. I therefore often use "level" and "reference frame" interchangeably.

⁷⁰ *Research: For AI, consider how one would evolve logic as a process over percepts in this manner. Consider what adaptive value this provides. Determine how a nonconceptual base can create the attribute of concepts.*

⁷¹ It might be prudent to find new terms rather than use "Percepts" and "Concepts" to avoid conflating with other ways they are used, but they seem intuitively appropriate for how I use them in this manuscript.

⁷² Others have also claimed that intentionality is not a fundamental property of percepts (e.g., (Ryle 1949)). However, I in general reject the externalist school of intentionality as grounds for my claims (Siewert 2017). Percepts should not be conflated with lower-level sensory processes. A percept, as I use the term, is the experience that accompanies grounding, including for abstract thoughts. It makes no claim on the "sophistication" of that experience. Nor do I contend that a nonintentional phenomenology is distinguished by being separable from intentionality. If intentionality is experienced, then it is necessarily a property of the given percept experienced. Percepts are not something distinct from concepts, but rather something that can include conceptual properties. Concepts, which are intentional, are a type of or part of a percept. They are important for higher level cognition but do not define it.

⁷³ (Block, 1995) proposes a distinction between Phenomenal and Access Consciousness. This appears to have many similarities to my distinction between Percepts and Concepts. There also appear to be some differences. The claims that follow could differ and be useful for both.

⁷⁴ *Research: Consider how this distinction for how concepts arise as percepts could be implemented in AI as a basis for cognition.*

⁷⁵ *Research: The Percept/Concept distinction may help gauge what separates us from other animals. Such a distinction might at least be a crude starting point for finer empirical tuning and extrapolation. I would perhaps suggest that all animals with nervous systems have perceptions, but only some have the conceptual type. Any animals incapable of conceptualization could not use explicit logic. They also could not be aware that they have perceptions. If true, then my claims suggest there is not aboutness for their perceptions nor self-awareness. Further consideration is required for what lacking logic and concepts entails. Another research question is what form a spectrum could take across animals, given this binary distinction between percepts and concepts. Are different animals limited in the types or degrees of concepts possible? How would conscious experience change given different limits to concepts? What sorts of properties must be available for percepts for different types or degrees of concepts to be available? I believe these are more interesting questions for understanding the "easy" problems on consciousness than simply trying to quantify it through activation and complexity measures. It may even offer some insight (but not proof) into what it is like to be different animal types, since we might be able to crudely consider what experience may be like with limited conceptual capabilities.*

⁷⁶ See for example: "Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference [or relation] to a content, direction toward an object (which is not to be understood as a reality), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired and so on. (Brentano [1874] 1973: 88)" (Siewert 2017).

at the level of a percept. The belief may be due to philosophy involving the contemplation of percepts, which turns them into concepts, thereby moving to a position outside the percept of interest to one about that percept or properties thereof.⁷⁷ A percept just is what it is. Within the Percept reference frame, it is neither true nor false and has no intentionality. To have an attribute of veridicality, the percept must be about something, in which case it is a percept of a percept. While percepts, at their own level, have no intentionality, logic, etc., if they include conceptual properties, then at the level of those conceptual properties, there is necessarily aboutness and logic. Once there is aboutness, there is logic associated with it, and there are therefore facts about the object that a perception is about. However, this is not the level of perception, which simply “is”.

Concepts are experienced just like all percepts. However, any experience, including those involving concepts, is always nonconceptual at the level in which it is experienced. Consider the following three cases: One can experience a percept that was evoked by seeing a dog. It has properties, but no concept is present. A concept would be to experience the thought that what one sees is a dog. The intentionality here is the perceptual property of directedness, as a process of conceptualization, towards the dog properties in this perception.⁷⁸ A meta-concept would be to experience thinking about the fact that you think it is a dog.⁷⁹ All three of these are still percepts, each different than the other. **They are experienced at some level that could itself be said to be non-conceptual, but the latter two have properties that contain concepts within that (non-conceptual) percept.**^{80,81}

⁷⁷ Some (but certainly not all) scientists and philosophers may be inclined to mistake concepts for the phenomena conceptualized. It is the nature of how intellectual activities are done to objectify phenomena; to hold them up as objects of study necessarily conceptualizes them. This enables analysis of the concepts of phenomena but not the phenomena themselves, at least not directly. The more nonconceptual richness there is for phenomena, the less complete are the insights and conclusions made about them through the analysis of their conceptual analog. Scientific theories on Music (e.g., what it is, why it evolved) provide a good example. The implicit essence can be difficult to evoke through our conceptualizations within the subject. Of course, only Philosophy can realize the distinction between concepts and the phenomena conceptualized.

⁷⁸ *Research: This distinction suggests a sort of “observer effect” that converts percepts to concepts. Although it is a stretch, consider if there is any relation to be made with the observer effect in Quantum Physics. One can say, phenomena cannot be described or measured as some “thing” at the percept level. Once described (i.e., “measured”), it becomes some “thing”, but it is now also a part of a new “pre-collapse” environment (Wigner 1961). Further consideration might include this consequence of the claim: There would always be a more fundamental “pre-collapse” within which any measured collapse is a property.*

⁷⁹ Percepts are not limited to direct perceptions and meta-perceptions. There can be meta-meta-perceptions (i.e., concepts about concepts), and so on, including self-awareness, where the self is conceptualized, and concepts about other’s thoughts (i.e., explicit theories of mind).

⁸⁰ A similar distinction seems to be made by Wittgenstein with proposition 3.263 of *Tractatus Logico Philosophicus* (Wittgenstein, 1922), where he describes simple vs. complex distinctions and proposes that the former can only be defined by use in the latter. This aligns with the claim that perceptions do not themselves have “aboutness” except when thought about (i.e., in the form of concepts). If this distinction is equivalent to Wittgenstein’s, then it follows from the claims above that complex distinctions belong to the set of simple distinctions, just as concepts are in the set of percepts.

⁸¹ *Research: Beyond Wittgenstein and possibly Block (see footnotes above), I expect that many others have made similar distinctions. I believe this distinction between percepts and concepts is critical to understanding ourselves and building AI. It would therefore be a valuable project to collect, compare, contrast, and integrate all insights*

Different concept meta levels can have different truth values. For example, a receiver of the message, “a dog is a cat” might experience the percept of “it is true that a statement was made”.⁸² This is a concept because it is an assertion about the non-conceptual percept of experiencing the receipt of the message. A receiver could also experience the percept of “this statement is false because a cat is a feline and a dog a canine”. This is a metaconcept. It is still about the statement received, but it is focused on the assertions it makes about the concepts of dogs and cats.⁸³

It is important to reiterate certain nuances in this Percept/Concept distinction. When I claim that percepts “just are”, I am making a claim about them at their own level. At this level, however, the percept can include intentional/logical properties, *not about themselves but about something contained within the experience of this percept*.⁸⁴ These intentional/logical properties are what I am referring to as “concepts”. Thus, percepts can be about percepts; they can have properties associated with perceived objects or events.⁸⁵

There is difficulty in explaining the percept/concept distinction due to the dependence on one’s frame of reference. There is a frame of reference in which any phenomenon is a percept. There is another frame of reference within which one can consider what the percept is about, but the percept experienced in this case differs from the percept being considered. Concepts are thus grounded as percepts (put more accurately, the perceptual experience of a concept is the phenomenological character associated with its grounding). This points to an issue with AI models that both model and manipulate concepts as symbols, which will be discussed throughout this manuscript.⁸⁶

and claims that have been applied to this distinction. For example, these should be compared to Merleau-Ponty’s (Merleau-Ponty, 1968) pre-theoretical layer, referred to as “perceptual faith”, upon which knowledge is grounded, as well as the phenomenology of Husserl (e.g., lifeworld) (Husserl, 1970), and Frege’s descriptions of Sense, Reference, Meaning, and Concepts (Frege 1892).

⁸² This example is admittedly awkward, but it readily bears its conceptual nature; one that is about what was experienced. It is conceptual because the experience that a statement was what was experienced is directed towards that perceptual experience, which could be true or false.

⁸³ *Research: Consider what would be required from a grounding perspective for AI to be said to itself form concepts. The “assertions” made in outputs are not necessarily intrinsic assertions.*

⁸⁴ Note that, since conceptualization is always about percepts but never the one within which the concepts are experienced, it can only ever conceptualize something extrinsic to the perception within which it is experienced.

⁸⁵ Note that not all language elements are necessarily true or false and therefore conceptual. Saying “hello” or “potato” may be best understood as nonconceptual. However, to claim any of these were said and/or how or what underlies them, is required to eliminate them. Also note that, if any such statements are taken to imply an assertion (e.g., “I am happy to see you” or “this is food”), then of course in that reference frame they are conceptual. As nonconcepts, they seem no more symbolic than seeing a certain color or texture.

⁸⁶ For one, I believe this distinction demonstrates the importance of generative capabilities for AI. There are infinite possible concepts. A non-generative approach to AI would be significantly limited. This includes any modeling approach that assumes explicit concept primitives, even if they are generative in a more limited sense. The set of concepts one might hope to explicitly model would be but an infinitesimal fraction. Consider that there are not only dogs and cats but also concepts of specific statements, thoughts about those statements, and so on. I believe this goes beyond recursion. Implicit combinations are possible, as well as contorting the qualities of different concepts (e.g., conceiving of a gigantic dog, neon pink cat, etc.). Explicitly modeling every possible concept at every possible meta level to produce a sort of hierarchical model of concepts would be an endless task. A more fruitful approach is

Let me briefly mention implications for consciousness. **Qualia are not inherently logical (i.e., at the level of the qualia) but instead can intrinsically have the qualities of logic (i.e., it is not a logical quale, but a quale of/with logic). Thus, Intentionality is not a fundamental property of consciousness. Only when there is a property of thinking about the object(s) perceived as an object(s) is there any "aboutness". Intentionality is a meta property for a percept. It is always intended toward something other than the percept within which it is contained. Pure conscious experiences = qualia = percepts, and percepts without meta-properties do not have "aboutness" and cannot be true/false.** This also means that physical phenomena, like all phenomena, can only be *conceptually* reduced to mechanisms (and even then, only in principle).⁸⁷

To claim that percepts can have meta-properties but not about themselves is to say that, just like intrinsic experience can have extrinsic experience properties and first-person experiences can have third-person properties, so too percepts can have concept properties. Indeed, the claims about percepts and concepts, as well as intrinsic and extrinsic emergence, align to the distinctions between first and third-person perspectives. **All third-person properties are extrinsic emergence, and all first-person properties are intrinsic emergence. All third-person properties are properties of a first-person experience. All first-person experiences just "are", while the third-person properties are about something. That is, percepts are first-person, while concepts are third-person. Concepts always occur with the property of a third-person perspective, even when about ourselves, and can in principle be fully eliminated. However, Concepts always occur within a perceptual base that is non-eliminable from the first-person perspective.**⁸⁸ **Intentionality is an extrinsic emergent property. The percept within which it exists thus has the property of being directed towards something but that something is not itself (i.e., the experience within which it exists). This makes it a third-person property.**

This has implications for how we cope with the world. Consider **affordances**, defined as what the environment makes available to (or through) an agent (Gibson 1979). The claims above entail that the *experience* of affordances does not require them being experienced explicitly as

to learn constraints within a continuous space with infinite possibilities, such that the constraints are built into how concepts relate and would be generated given specific stimuli and current model states. This allows for common concepts to be identified while also ensuring novelty of what is experienced is also captured. For example, the location of the concept for dog may be a subregion and every time a specific point in the subregion is identified differs slightly (perhaps an attractor gets one closer to the prototypical concept in most cases, or perhaps there are two separate spaces; one with an attractor to the prototypical concepts given stimuli, and one that captures what is episodic/specific to each given message that would evoke a concept). I say more about this in later chapters.

⁸⁷ *Research: Again, there may be important theoretical work in understanding what if any implications this has for Science and Knowledge. Discussions are experienced but what is experienced is never the same as the experience discussed (i.e., the experience that the discussion is about). If experience and knowledge are equivalent, then what does this distinction mean when discussing (internally or externally) fundamental physical entities and processes? Does it matter? Why or why not?*

⁸⁸ *Research: Consider how these claims relate to C.S. Peirce's concepts of firstness, secondness, and thirdness (Deledalle 2000). Perhaps secondness and thirdness, as one basic type that could be thought of as a meta-firstness, are both special types of firstness, and always include firstness qualities. Further research is needed to determine if there is a difference unaccounted for at my level of claims, with respect to AI, the Mind, and Semantics.*

affordances. Affordances, in their most fundamental form, implicitly shape percepts. We can certainly conceptualize affordances, but this is derivative of their implicit influence on our experiences. Thus, **if we assume AGI requires an understanding of the environment in terms of affordances, then it is entirely insufficient to simply map objects in the agent's world to affordance types. Affordances are most naturally understood as constraints on percepts, rather than conceptual properties. An agent may be able to conceptualize affordances, but the means to do so should not be conflated with the fundamental agent-environment processes from which we would be externally defining (i.e., conceptualizing) phenomena as affordances.**

Affordances simply offer one example of the potential to misrepresent phenomena in AI modeling. More generally, **the fact that we must reason at a conceptual level to understand and discuss nonconceptual content may be at the core of why we are challenged to explain the Mind in relation to AI. This may also relate to why we might be inclined to assume intentionality is a property of all conscious percepts, even nonconceptual ones.**

The aboutness of something perceived differs depending on the emergent level of the properties in the percept. For example, a car, like anything, can be reduced to a collection of atoms and forces. However, a concept for the car is typically intended towards perceptual properties that may be evoked by the underlying atoms and forces but are nothing like them. Instead, the perceptual properties intended toward are emergent (i.e., heuristic). The experience is intrinsic emergent, and can be conceptualized as intending toward (i.e., emergent-ally extrinsic to) something we describe as a car.

And yet, the same degree of derived intentionality would be true at any emergent level. There is an intrinsic experience underlying the intentionality toward any concepts. A question I wish to raise here is whether this holds true even of concepts for the lowest level mechanisms.⁸⁹

First, it is worth distinguishing intrinsic vs. extrinsic emergence “levels”. I mentioned earlier that emergence may not be fundamentally hierarchical. More specifically, it may be a mistake to fundamentally describe *intrinsic* emergence as hierarchical. **However, Concepts are clearly structured in hierarchical form.** For example, a number of individual experiences can be grouped into the concept of a cat, individual cats into the species Cat, the cat species into the genus, Felis, etc. This is of course well-known, and includes the other hierarchical elements and levels, such as Physics to Chemistry to Biology, etc.⁹⁰ As I mentioned, this hierarchy may itself be a conceptual

⁸⁹ Research: There is something worth considering here that does differ between the intentionality of something like a car and something like an atom, which is that the former is directly experienced. What are the implications for things that are only ever experienced as concepts, such as atoms? It may be fair to say that any phenomena that are only ever experienced as conceptual are not emergent in and of themselves; only the concept of them has intrinsic emergent properties. In other words, there is no way to intrinsically experience the phenomena themselves. We only know of them through conceptual experience. If so, then mechanisms are not emergent because they are never intrinsically experienced. I make this claim later but mention it here as something that deserves further consideration.

⁹⁰ Research: Note that we thus have two conceptual hierarchies: One for the meta levels of percepts (e.g., a movie vs. a movie about a movie, etc.) and one that orders concepts within a meta percept level (e.g., a movie vs a story vs. entertainment). Are these orthogonal hierarchies or could one include the other? Are felines a concept about individual cats? Are mammals a concept about felines, but also other animal types? How could the following be explained in terms of sets and members: “I know that you know that I know...”?

emergent construct. There is nothing absolute, complete, or fixed about it. However, concepts are perceived as hierarchically structured.⁹¹

Note that there are infinite possible concepts at each level that could be related to a percept. For example, above the concept of cat could be *Felis*, the set of cats in my living room, etc. This makes fully modeling the conceptual universe effectively impossible and relates it to the Frame Problem in AI (Shanahan 2016).⁹²

There are interesting limits for what can be perceived in this conceptual hierarchy. Why is the concept of mammal often accompanied by the percept of a specific example of a mammal? I can think of two possibilities. The first is that one can only perceive of concepts that are single member sets. The other is that it is a contingent matter of our evolutionary history that percepts are comprehended in terms of sensory experience, and sensory experience is rooted in the physical environment. These both relate to the fact that anything physical is an instance of something specific in space and time (e.g., only one specific cat can be in the exact same place and the same time).⁹³ However, these possibilities raise other questions. For example, how can we comprehend higher level concepts if we cannot perceive (i.e., ground) them? We clearly can, and I will suggest later that comprehension may be its own type of sense.

There are issues with the claim that all perceived concepts are single member sets. We clearly cannot experience the full number of attributes of a given concept, but any attribute left unperceived opens a lower level of possibly infinite members. For example, if I see a cat but my perception of it does not include the number of hairs on its belly, then there is an essentially arbitrary number of cats that are members of the cat concept I perceive; each with a different number of hairs on its belly.

Perhaps there are interesting questions to pursue here, but the point is raised simply to demonstrate that there is something wrong with any general approach that attempts to explicitly model percepts into a hierarchical model.⁹⁴ The percepts just “are”. They are not *fundamentally* hierarchical. Only the conceptual perspective is hierarchical. The confusion disappears when we see that the grounding of concepts relates to how concepts are perceived. So, when I conceptualize a mammal, I can reliably conceptualize it as a higher-level concept than any individual mammal. However, the grounding of the concept of a mammal (i.e., the way this concept is perceived) has no fundamental hierarchy. It just is. What this all suggests is that **modeling knowledge as a hierarchy, while (potentially) consistent as a universal account of**

⁹¹ Note that the core relationship between Percepts and Concepts is different than relationships with respect to what have been called “Semantic Primes” (Goddard and Wierzbicka 2014). The relationship also differs from Composites, or Basic vs. Abstract Domains (Green and Evans 2006). A percept is not distinguished by being simpler or as a building block for concepts.

⁹² *Research: Does this mean that the hierarchy of sciences is also necessarily incomplete? Are the concepts we have constructed logically arbitrary but selected because they are effective?*

⁹³ *Research: Consider how this may relate to Plato’s Ideal Forms (Theory_of_forms n.d.), which can be conceptualized, but only imperfect instances can be perceived.*

⁹⁴ This does not mean that AI models should not have hierarchical architectures at subsymbolic levels, thereby enabling hierarchical properties to be grounded in the given agent’s understanding of the world. This claim here is instead that breaking down concepts of percepts explicitly into hierarchies does not ground those percepts.

concepts, will not itself ground the concepts. This alone would not provide any means for an agent to comprehend any of the concepts modeled for it. The hierarchical level of a concept is best understood as a potential property of the perception containing it. It is derived from, rather than fundamental to, the percept. We will see that the value of these conceptual properties is that they can be used as symbols, which can be communicated to evoke the percepts. This is Symbol Grounding.

This all should either seem odd or of little importance for AI. Consider the relationship to Deep Neural Networks (DNN) that capture increasingly abstract properties from data across neural layers. If these DNNs could experience anything, it seems that they could equally experience across the conceptual hierarchy. However, I do not believe this contradicts the claims in this section. While the percepts that DNN's might experience can be described as hierarchical in the sense of being of increased abstraction, such a claim represents them *as* concepts. Still, this seems like a cheap resolution, given that any claim is conceptual. It is true that the first-“person” experience of the DNN is non-hierarchical, and it could be further argued that the DNN's percepts could in theory include conceptual properties (i.e., of other percepts). However, as an external observer, it seems more than simply practical to claim that the features processed are hierarchical. Some percepts are clearly more abstract relative to other percepts. However, I think this just demonstrates how critical concepts are to our understanding of the world. The point in this chapter is emphasized to clarify the difference between phenomena in their percept and concept forms. This distinction should not be seen as diminishing the importance of concepts. Without concepts, there is no shared knowledge. **We communicate in Concepts (Indeed, this is why it is so difficult explain Percepts).**⁹⁵ Science itself is conceptual. But the Mind realizes any concepts as properties of a more fundamental perceptual base.⁹⁶

I claimed above that there is an alignment between Percepts/Concepts and First-/Third- Person Perspectives. There are also other seeming dichotomies that are each more appropriately understood by distinguishing a fundamental phenomenon vs. a subset or property of a fundamental phenomenon. Specifically, all Objectivity (i.e., objective knowledge and objective reasoning) is a special case of Subjectivity, and our Public world is a special attribute of our Private world. These claims can be put together in the following way: Our cognitive processes are fundamentally percept-based, which is subjective and private. Concepts, Objectivity, and Public phenomena are only represented as such from this base. I would guess that the most

⁹⁵ This may not be universally true. There are speech acts that seem to fail to make explicit assertions, such as “Ouch” or “Hello”. I will generally be restricting what I claim about thinking and communication to inferential thoughts and statements. These are sufficient to meet the claims in this manuscript, including the claim that scientific and philosophical communications are conceptual.

⁹⁶ *Research: Consider whether concept properties in conscious experiences are always about the past, such that a present experience (that actually experienced experience) must be perceptual (and, in the reverse direction, percepts must always be in the present). If true, then this is another distinction between concepts and percepts. Concepts are not directed toward what is experienced in the present within which they are experienced, while percepts only exist as experienced, which entails them being in percept-form only in the present. If this is true, then there may be experiments that could distinguish them in principle. Perhaps conceptual properties have some dependence on memory but nonconceptual percepts may not (or at least different requirements).*

controversial of these claims is the one for Objectivity. Everything we (validly) categorize as objective is completely reliable and as precise as the claims allow. $1 + 2 = 3$ is always true (i.e., reliable), and it equals exactly 3 every time (precise). This precision and reliability, I believe, relate to what Eugene Wigner's described as "the unreasonable effectiveness of mathematics for the natural sciences" (Wigner 1960).⁹⁷ I consider this unreasonable effectiveness an unsolved problem. The challenge to my claim that objectivity is a special case of subjectivity is that we can use objective processes (e.g., sophisticated math) to arrive at answers we have never empirically experienced that turn out to be reliable and precise when finally applied or observed. Also, a need for precision and reliability for evolutionary fitness could be a necessary means to reach objective knowledge, but it is unclear that it would be sufficient; we seem to be leveraging something universal about reality when we use math/logic. I will simply note here that, even if it is something fundamentally built into the universe (or something that builds the universe), it is not a given that it is the same form in our own percepts as it is out there. Mathematics might be a tool that helps cope with the world, such as by breaking it into sets, without it existing Platonically. Still, this leaves open the question of what it is exactly, and how it came to be if somehow not fundamental.⁹⁸

Of course, all Science is conceptual in nature. Thus, all our models of our universe, across the sciences, are concepts. Concepts are critical. Moreover, even if they cannot explain perceptual experience, concepts can certainly evoke them.⁹⁹ If essentially the whole formal enterprise of knowledge is composed of Concepts, then why does this distinction between them and other Percepts matter? Concepts are *about* the nonconceptual phenomena; they are not *the* nonconceptual phenomena. I am claiming that the process of eliminating extrinsic emergent properties involves the conceptualization of those properties, and, as I have already claimed, intrinsic emergence itself cannot be eliminated. This implies that **intrinsic emergent processes cannot be fully conceptualized. Thus, concepts cannot themselves explain how they are evoked. Without modeling how Concepts are generated, we cannot hope to build AGI capable of what we do.** Models built from Concepts are missing the implicit processes that determine how the Concepts would be evoked. These implicit properties are essential for grounding concepts. They are what drive how concepts are evoked and utilized. They are also the perceptual properties associated with our concepts. My claim is that these implicit processes matter for

⁹⁷ Of course, that $1+2$ always equals 3 is perhaps the least interesting case of unreasonable effectiveness. Wigner presumably is more concerned with how abstract and often arcane mathematical derivations performed without any consideration for their consistency with the universe can end up precisely describing physical processes. $1+2=3$ might be seen as the reverse situation: how do complex physical processes always end up consistent with the same basic mathematical assertions?

⁹⁸ Candidate approaches to addressing this are planned for a future manuscript. In short, these involve reducing analytical procedures to evolutionarily plausible primitives in the form of referencing actions.

⁹⁹ *Research: Are Logic and Math always conceptual? Are they best understood as a tool(s) for conceptualization? If so, and if the universe, as we are able to know it, were fundamentally mathematical, then what are the epistemic and ontic implications? Are they better understood as conceptual building blocks? Are atomic elements tools rather than ontologically fundamental? Alternatively, are atomic elements and fundamental forces experienced with qualia properties that are not true or false?*

Intelligence. This **grounding is not merely epiphenomenal**.¹⁰⁰ There is an asymmetric relationship between Concepts and Percepts, such that, to attempt to emerge Grounding by starting with Concepts is backwards and cannot work. The nonconceptual properties provide the **grounding** for any concepts, where the concepts stand in as symbols for some phenomena. It could be argued that **the very function of concepts is to evoke the grounding that ultimately helps us cope with the world. In turn, the grounding evokes concepts for them to serve their role, which is performed via communication and explicit thinking**.¹⁰¹ As I will explain in a later chapter, without grounding, we may be able to build models of domains, but not modelers. It is therefore critical to determine how these implicit processes must themselves be modeled in order to apply them effectively, including the dynamical processes by which they construct Concepts. This role for grounding, and how it might be appropriately modeled, is discussed throughout this manuscript.^{102,103}

Complexity, Emergence, and Consciousness

How does Emergence arise? The most common explanation seems to be that emergence forms from complexity. **While properties associated with complexity may be required for (or perhaps in some way proportional to) emergence, nothing about emergence is explained by claiming that it arises from complexity. The properties associated with complexity (i.e., interaction, interdependence, adaptation, etc.) provide no insight into emergence itself. For complexity to be an explanation of emergence requires an understanding for how the mechanisms emerge the global properties. Otherwise, this merely demonstrates that one is always present with the other.** Referencing an association lacks explanatory power.

Moreover, **the novel properties we observe arising from complex systems are examples of extrinsic emergence. Since extrinsic emergence is a property of intrinsic emergence, and is eliminable, we are ultimately interested in how intrinsic emergence, which is the same as consciousness, arises from complexity.**

Since we project extrinsic emergent properties (onto external phenomena) as an intrinsic experience, consciousness explains emergence, not the other way around. More precisely, Consciousness (i.e., The Mind) is not Emergent, it is Emergence.

¹⁰⁰ Even if Grounding was epiphenomenal, we are stuck using the associated abstract heuristic concepts for explaining reality due to the Description Gap.

¹⁰¹ The nonconceptual base used by mathematicians and logicians to ground numbers, operators, etc. may be how the mind achieves novel mathematical and logical insights, where the process of arriving at insights lacks explicit mathematical or logical explanation.

¹⁰² *Research: Consider if Concepts as tools could be innate rather than learned. Would this distinction help explain why some structure seems innate (e.g., aspects of language (Chomsky 1990))?* Does it explain why proposed innate structures are often violated?

¹⁰³ Understanding how thinking about and discussing percepts converts them to concepts, and how concepts are subject to eliminative materialism, might be relevant to various open philosophical questions and paradoxes. Although mostly beyond the scope of this manuscript, it seems worth reflecting on this process when considering past and future philosophical arguments. I do not claim this will resolve all such problems, but it may be for some that conflicts are due to not recognizing the differences between the perceptual and conceptual forms of phenomena of interest.

Although commonly stated (e.g., (Page 2009, Feinberg and Mallatt 2020)), there are issues with claiming Consciousness is an example of an emergent system. It is intuitive to suggest it emerges from complex physical process (i.e., brains). It is of course a property of what we project onto other systems (e.g., other humans). However, since this approach focuses on extrinsic emergence, it is open to eliminative materialism. My point here is not show that the brain cannot emerge consciousness but to bring to light the role of third- vs. first-person perspectives (i.e., concepts vs. percepts) in the challenge of explaining emergence from physical properties, regardless of their complexity. These distinctions, discussed at length above this section, appear to have an unappreciated role in the so-called, Mind-Body Gap.¹⁰⁴ The belief that emergence explains consciousness arises when we confuse our extrinsic emergence as something else's intrinsic emergence. Complex Systems certainly emerge novel properties, but emergent features in the world are emergent only to a conscious perceiver. **It is therefore uninformative to claim that consciousness is like other emergent properties.** Yet these comparisons are common in scientific literature.

Consider Functionalism, which claims that it is what computations do that matters and not the medium upon which computation is performed. A typical analogy is that the mind is to the brain what software is to the computer. However, assuming the mind is like the computer software merely points to an example of the very mystery associated with the Hard Problem of Consciousness: The emergent properties that constitute the software only exist in the mind of the observer. We project our conscious understanding of what the computations do. That is, the functions are projected onto the external systems. Thus, **software is not to the hardware as the mind is to the brain. Properties evoked by computers are not *intrinsically* things in the world but things in our mind. Software properties are constructed by an observer and projected as concepts in the world. They are not things in themselves separate from the mind. Similarly, mental emergence cannot be compared to how the properties of water emerge from H₂O molecules (another common analogy). The mystery of how H₂O emerges properties of water, being emergent properties intrinsic to the observer, is the very mystery we are trying to solve for consciousness.** Again, if there is any confusion, it is because the mental projection about water happens to have the property of being about something external to us. If complexity is a prerequisite for consciousness (intrinsic emergence), then the real question is *how* does complexity give rise to mental percepts and concepts? What is it specifically about complex systems that would explain conscious experience? Why does emergence occur at all?

One might argue that consciousness arises by having an emergent representation internally of our own minds. However, this either pushes the problem back one step or creates an ever-

¹⁰⁴ Note that the Intrinsic vs. Extrinsic Emergence distinction differs from the common distinction between what are called Strong and Weak Emergence (Bedau 1997), where the emergence of the Mind is in the Strong category as an example of something not in principle explainable from its parts. While interesting in its own right, I do not think the Strong vs. Weak distinction gets to the heart of the disconnect in understanding Mind or Emergence. Extrinsic Emergence is special case of Intrinsic Emergence. They therefore have an ontological order. The Strong vs. Weak Emergence distinction provides no insight into the fundamental standing of Mind as *Emergence*, rather than as *emergent*.

oscillating conflict, because it means the emergence begets the mind, while the mind begets emergence. Instead, Emergence *is* mind and mind *is* emergence, and there simply is no underlying lower-level explanation of the Hard Problem; it is always that which is different from its mechanisms (different than the sum of its parts). This distinction does not resolve the paradoxes, but it seems critical to framing the discussion. Fortunately, I believe we may be able to make progress for AGI without resolving this problem. In the above section, I described how grounding is not epiphenomenal. I will suggest ways to implement grounding in later chapters.

Emergence and Mechanisms

So far in this chapter, I have not included lowest level mechanisms within the set of emergent phenomena. If all mental experience is emergent, then it would seem that intrinsic emergence must cover all experienced forms of phenomena, including for mechanisms. All experience is of emergent properties, even the experienced property of external physical things.

Before moving further, recall that claim that there is no fundamental hierarchy for intrinsic emergence phenomena, but there is an intrinsically perceived hierarchy for concepts.¹⁰⁵ Thus, when we move down the perceived conceptual hierarchy, percepts do not follow. From an extrinsic emergent perspective, this leads to their elimination. However, intrinsically, they are not eliminated any more than changing the subject in a discussion eliminates what was discussed prior. It is perhaps appropriate to say lost percepts are not associated¹⁰⁶ with the mechanisms, just as percepts change when someone changes their view, what they are listening to, etc. This, I believe, is a way of understanding, in part, why emergence is novel. As one moves through the hierarchy, intrinsic experience changes, but not in a way that is explained by lower levels.¹⁰⁷

What this means is that there is no basis for any intrinsic emergent phenomena to be fundamentally of lower order. Even though mechanisms are conceptually fundamental, the concepts for them can be grounded like anything else without there being an infinite regress for

¹⁰⁵ The perceived hierarchy for concepts offers no means to describe (i.e., explicitly map) between hierarchical concept levels due to the Description Gap, but it is the hierarchy upon which eliminative events might be perceived to occur. Note that this hierarchy is a construct. It may be incomplete and can change over time. For example, in principle, some new science could eventually be born between physics and chemistry, or some species may be split into two separate species.

¹⁰⁶ This association is best understood as Symbol Grounding, which is discussed in depth in the Information chapter.

¹⁰⁷ Keep in mind that Eliminative Materialism contends that emergent phenomena are eliminated given underlying mechanisms. They are not reduced to them, as if the lower levels offer an explanation for the emergent properties. I have claimed that this is the proper description for extrinsic emergence. I have also claimed that a consequence is that concepts are lost, including functional explanations. Answers sought at the emergent level cannot be satisfied through elimination. Emergent concepts are not explained by mechanisms; they are eliminated. However, because there is general asymmetry in formal causal relations between conceptual levels, they can be conceived as hierarchical. But doesn't a hierarchy entail a reduction? If so, this would seem inconsistent with the claim that the lower-level concepts eliminate the higher-level ones. Consistency is regained by understanding that this hierarchy is a construct that places concepts in relation to each other but does not explain how the concepts constitute each other. In other words, the Description Gap described earlier prevents any possible reduction or explanation. We simply know there is an eliminative association between something at a higher level and a lower level (knowledge of elimination as a process between two concepts necessarily exposes an association).

what mechanisms then emerge that grounding.¹⁰⁸ Since we can understand mechanisms, this understanding must be enabled through nonconceptual properties.

Mechanisms are intrinsically perceived, just like all other phenomena. However, they could be described as being perceived as non-emergent.¹⁰⁹ This is the nature of the hierarchy being itself a concept. If mechanisms were non-concepts, then their perceptual properties would be the same as their conceptual properties. However, if concepts require nonconceptual phenomena to be perceived, then we would be unable to intrinsically conceptualize such mechanisms. Of course, even concepts are pure as mathematical and logical symbols can evoke a variety of nonconceptual properties for different people. There would thus be a disconnect in claiming they are *conceived* as nonemergent. However, claiming that they should be considered as having the property of being non-emergent is more reasonably a claim about the mechanisms “out there” not our concepts of them. I see no issue with such a claim. In fact, if there is anything outside the mind, then it is necessarily only non-emergent mechanisms, given my claim the Emergence is equivalent to the Mind. This simply requires that such mechanisms are not conflated with our concepts for them, which are emergent, since they must be grounded to be known.

Thus, this is a claim about mechanisms as concepts, not physical processes “out there”, and it is important in considering Knowledge (and therefore AI). As concepts, a mechanism-only *explanation* provides no means to emerge the perception of water from H₂O-level properties. That is, it cannot account for the difference in intrinsically perceived emergence across conceptual levels. The same holds for mapping wavelengths to colors and any other attempts to map across concepts that differ in their non-conceptual properties. This goes beyond mechanisms and applies to any representation of knowledge that consists of only concepts with no grounding. It suggests **a model with only concepts and no grounding has no means to emerge grounding, and that grounding must be nonconceptual.**

We have finally reached tangible implications for AI, so I will reiterate this point. Grounding is not accomplished by just mapping the emergent *concepts* to mechanism concepts (or between any concepts). This does not entail that mechanisms themselves, as opposed to concepts of them, cannot ground other concepts. What it does mean is that, if mechanisms were capable of grounding concepts, they themselves could not be *modeled* only as concepts. We will see that this requires criteria for modeling them subsymbolically. Moreover, given the role of nonconceptual properties, what grounded emergent concepts are evoked depends on the receiving agent’s properties. Mechanisms are only sufficient conditionally on the receiving agent. All of this means that for AI, grounding involves more than simply modeling a mapping from

¹⁰⁸ That is, there is no issue in our concepts for mechanisms being grounded in e.g., neural processes. We will see that this requires that those neural processes are subsymbolic to the concepts they ground, including for mechanisms.

¹⁰⁹ This claim holds for logical mechanisms. Analytic statements must have nonconceptual properties to be experienced at all, but they can still be conceptualized *as having the property of* having no nonconceptual properties. The difference is a matter of Reference Frames, which are discussed later.

higher-level concepts to lower-level concepts.¹¹⁰ The AI affects the grounding, and thus what any concepts mean.

From Intrinsic Emergence to Grounding

As a reminder, Emergence (mind) is a functional heuristic in that it helps us act in the world.¹¹¹ When sensing parts of our world, humans (or any living organisms) cannot sense or process all of the underlying mechanistic properties and their interactions. Yet, we need some way to cope with phenomena that would otherwise have prevented our ancestors from passing on their genes. Our nervous system has adapted to construct and process heuristic features. The associated stimuli are received, and the brain constructs effectively reliable signs (e.g., symbols) to associate with the phenomena we select them to reference. For example, I cannot see the individual molecules that together constitute a snake, but my brain creates an emergent heuristic form that is effective from an evolutionary standpoint to cope with them.

From an Information Theoretic perspective, these heuristic representations (e.g., symbols) can be considered compressed, often lossy, representations for the underlying phenomena. More accurately, they model what is evolutionarily relevant¹¹² given the underlying mechanisms. From an extrinsic perspective, these models are adaptive rather than accurate representations of reality (Hoffman 2019). These are what observers mentally project onto mechanisms in the world for which we do not have direct access (Rorty 1979); they do not exist “out there”.

In some ways, our ignorance of reality is like a computer’s ignorance beyond its program. We perform processes that result in models of something out there, but we have no obvious way to gain understanding of what it models. Likewise, external modelers may know what a computer model represents, but the computer does not. Ultimately, all we know about reality is how effective our representations are, not how accurate. There is a regress here, however, as any concept of effectiveness is itself a model of some inaccessible reality.

In summary, Intrinsic Emergent experiences are shaped by how effective they are as heuristics, whose function it is to help us cope with our inability to fully measure and process underlying mechanisms of systems. Intrinsic Emergence has a functional role in constructing compressed representations. However, any functionality should be amenable to scientific explanation. As explained earlier, the Sciences build concepts; hypotheses and theories only cover extrinsic emergence, which is consistent with Eliminative Materialism, at least in principle.¹¹³ Moreover, as claimed in the AI-Automata thought experiment above, AI heuristics cannot be claimed to

¹¹⁰ Again, mechanisms *as concepts* cannot ground higher-level concepts. We will see that anything can be nonconceptual relative to other reference frames, and this includes mechanisms. This simply requires that the mechanisms are not symbols from the given reference frame. As I write this, I can assume there are physical processes that evoke my grounded thoughts and the letters typed on the screen. Those underlying processes are not symbolized/conceptualized in the grounded emergent phenomena they evoke.

¹¹¹ This claim is itself emergent. Unlike the similar challenge for eliminative materialism (i.e., that it eliminates itself), there is no contradiction in emerging the concept of emergence (or function). Still, the Description Gap leaves us without any explanation for how this comes about.

¹¹² Actually, it seems even more accurate to say they model dynamics “in a way that is evolutionary relevant” rather than “what is evolutionary relevant”.

¹¹³ This may explain why most scientists operate as functionalists.

demonstrate intrinsic emergence for those AI agents/models. This seemingly creates a dilemma. Extrinsically, there is no functional role for emergence since it can be eliminated, but to ascribe functions to intrinsic emergent phenomena is to conceptualize them. How can intrinsic emergence be both functionally effective and outside of direct scientific study?

I believe the answer is relatively straightforward: It is a *concept* of intrinsic emergence (i.e., qualia) that it is perceived as functional. Functionality itself is emergent and in principle eliminable. There is no issue here. As already discussed, it is the nature of Science and Philosophy to apply emergent concepts, even if those concepts are eliminable in principle. Scientific concepts are intrinsically experienced as external phenomena. While the functions that come with this intrinsic emergence as a compression process can be explained without any functional role for qualia's phenomenological character, the nature of describing it at a heuristic level entails the assumption of intrinsic emergent phenomena. That is, the representations are used in the process of thinking and communicating about their own nature. There is a relationship here to first- vs. third-person perspectives as discussed earlier.

To put this another way, intrinsic emergent phenomena, including extrinsic emergent concepts, are functional within a particular reference frame, where the emergent concept of functionality exists. This is a reference frame we are used to, where entities are experienced as having functionality. The concepts are eliminable in theory, but so too is the concept of functionality itself. It thus only makes sense to even talk about heuristics at this emergent level.

However, at this level, it also makes sense to discuss this functional form as grounding. At an emergent level, where functionality exists as an emergent concept, grounding could be described as the nonconceptual experience. Regardless, intrinsic emergence itself, as the foundation of concepts, is simply too confusing to use as a concept.¹¹⁴ This is the rationale for why I focus on grounding, which can be viewed as the extrinsic form of intrinsic emergence. **Grounding does not provide any insight into the phenomenal character of intrinsic emergence, but it allows us to describe, and even define it in its extrinsic form, as we will see in later chapters. The extrinsic properties of intrinsic emergent experiences function as heuristics. As Science is *about* phenomena, it is appropriate to focus on the heuristic properties of extrinsic emergence.**

Emergence and Science

Thus, emergence encompasses everything except we know. As already discussed, even though Science is a process directed towards modeling fundamental mechanisms, it is our process of understanding the world through the emergent concepts it constructs along the way. Any understanding it builds is adapted and colored by one's evolutionary history. Features in data, and the insights provided through analyses, represent emergent symbols upon which statistics and logic act. In fact, Statistics itself assumes an emergent concept; namely, a system. Isolated

¹¹⁴ Unlike Grounding and Intrinsic Emergence, "Percept" and "Emergence" are more entangled in common use, so I apply them more freely (i.e., functionally as equivalent to grounding and phenomenologically as equivalent to qualia and intrinsic emergence). It would be simpler to have avoided them altogether or arbitrarily place them on a side, but it seemed better at this stage to connect the claims in this manuscript with well-known (i.e., semantically rich) terms.

entities have no intrinsic statistical properties until they are related to other entities.^{115,116,117} Even aggregates of objects are composed as a set of some things. I believe eliminative materialism does not reach its end stage until it no longer makes sense to discuss or perceive of the mechanisms remaining as existing in some larger set or system. Elimination and/or Reduction is incomplete until there is only the “perspective” from the fundamental elements.¹¹⁸

Thus, anything described at a statistical level is emergent. The moment we describe any phenomenon with statistics, we are describing extrinsic emergent phenomena. Science relies on statistical concepts, and it is therefore important to keep this in mind. While this is obvious in and of itself, it is the claim that any emergence is projected onto systems from the observer that makes this potentially controversial. Likewise, **anything emergent, in principle, has statistical properties (in their conceptual form)**.¹¹⁹ This fact reinforces my earlier claims that emergent heuristics could be evaluated in principle in terms of loss compared to their underlying mechanisms. Concepts tend to be increasingly lossy with respect to underlying mechanisms as one goes up to higher levels (whether the hierarchy is itself real or a heuristic concept).¹²⁰

¹¹⁵ I am assuming a classical system here. A claim on whether quantum-based probability is emergent is way out of scope for this manuscript.

¹¹⁶ They can also have statistical properties when related to themselves at previous timesteps. However, such an analysis still forms a system of entities with different time values, whether or not the observer grounds them as a single entity.

¹¹⁷ This is not claiming that an observer would not observe any such entity statistically. Ignoring entities they interact with would surely create stochastic effects that would force a statistical understanding for the isolated entity. I am making a claim here about the entity itself, not how it is modeled from observation. A simpler way to say this is that there are no entities “out there” that are the set of some things in addition to the things themselves. Thus, to have a notion of a set of things requires an observer to conceptualize it.

¹¹⁸ or events, or points along a field, or whatever makes up the ultimate primitives.

¹¹⁹ This may explain why entropy and information seem so fundamental. All emergent *concepts* can in principle be characterized using Information Theory. However, concepts require some observer to form them. The Information chapter elaborates on this claim.

¹²⁰ *Research: Consider new ways to measure Emergence. Here are some (somewhat random) thoughts:*

- *To gain insight into emergence through models, analyze the similarity or relationship/function between dynamics of neural networks representing Reinforcement Learning agents and the environment they perceive (after effective and semantic processes have evolved, discussed in a later chapter). Is there a way to assess what static emergent objects “look like” for those neural networks vs. in the environment? Or must emergent objects be understood as dynamic processes? How do they differ from humans monitored in those same environments?*
- *Evolve agents in a virtual environment with tasks that require different degrees of coordination. Then, measure what level of scale vs. variety (Bar-Yam 2004) is adapted for different tasks based on some correlation measure(s). If scale here is truly equivalent to emergent scaling, then does this provide a way to measure it, along with its functional value?*
- *Consider how one might architect a grounded causal network for a resolution/scale-based parent-child hierarchy rather than parent-child relations. This captures Formal rather than Efficient Causes (Four causes n.d.) (or formal conditions). However, it might be interesting to discover what efficient causes drive formal causal relationships.*
- *Determine if it is possible to develop a metric for Markovian Distance, where data are measured for how far away from Markovian they are (or how much complexity is required to make them Markovian). Depending on comparisons to ground truth, the measure could potentially be used to determine how emergent or lossy*

If Science, and essentially all explicit knowledge, deals with emergent phenomena, then this should surely affect how we view and even do Science. These heuristics have been primarily shaped by how they help us cope with our environments. However, this is particularly important for AI, where we are generating systems to also cope with environments and need them to learn and exploit heuristics to deal with intractable problems, which are the rule rather than the exception. While Emergence has been studied for well over a century, I believe the points made above, including its all-encompassing nature, warrant increased focus with respect to how we generate AI, and its implications across the Sciences. What are the implications for emergence being mentally projected onto systems? Does it allow effective use of our collected knowledge about these emergent structures beyond a focus on what they reduce to or why they do not reduce to it? Or, if it is emergent all the way down, does it change nothing? How does the lack of an explicit map between emergent levels influence how we link knowledge across scales?

As mentioned above, heuristic representations are shaped by their effectiveness for ancestors passing on their genes. This provides a relationship to help define three related areas of Information Theory raised, but not formally defined, in (Weaver 1953). The relationship between emergent heuristics and the different types of “Information” Theory is the focus of Chapter 4. The next chapter helps set the stage by explaining the importance of performance (or effectiveness) in Intelligence, which has been relatively underappreciated in AI’s recent history.

data are with respect to the "true" classical properties of interest, which are assumed to be Markovian. This may not ultimately measure lossiness but might be an interesting measure itself.

Chapter 2: Performative Systems

Potential Limits of Representation-Driven Learning

In ‘The Cybernetic Brain’ (Pickering 2010), Andrew Pickering uses the term ‘Performative’ to set apart Cybernetics research from what evolved as AI. Cybernetics was a burgeoning field in the mid-19th century that focused on how feedback governed dynamics for different systems. Understanding the brain was a major application of the research performed. In the 1950’s, AI essentially split from Cybernetics with a focus on Reasoning (particularly in the form of symbol manipulation). These were two quite different ways of understanding Intelligence.

A canonical example is the research on homeostats (Homeostat n.d.), developed by cybernetics pioneer, William Ross Ashby. Although cybernetic research was important in advancing predictive methods, Prediction does not seem to be the central objective with these systems. Instead, the way these devices adapted to environment conditions maintained their homeostasis. Given that a core driver of living systems is to maintain homeostasis, it can be argued that the function of intelligence is primarily for this purpose.¹²¹ Even sophisticated long-term planning could be viewed as maximizing homeostasis over longer time scales. Rather than reasoning with symbols, the “intelligence” of the homeostat was its ability to cope.¹²²

Since the split between Cybernetics and what has been called “Symbolic AI” or “Classical AI”,¹²³ it is that latter that is generally assumed to be the forbearers of current AI, even though Machine Learning owes perhaps more to Cybernetics with its utilization of feedback loops to adapt or learn. While some AI areas have a performative emphasis, particularly in Reinforcement Learning, there remains a general assumption that veridical representation is the path to Intelligence. In contrast, there is a lack of appreciation for assessing models, representations (or a lack thereof), etc. by their role in agent effectiveness in the world. While it may be a crude interpretation, I see the split between Cyberneticists and what became AI as a contingency, where something core to advancing AI was lost, including knowledge of the general philosophies that drove the spirit of Cybernetic research. I therefore borrow the term “Performative” to retain the overall features distinguished by Pickering, even if what I describe as ‘Performative’ diverges beyond what one might consider the domain of Cybernetics.

Setting Cybernetics aside, since the beginning of AI proper, there have been strong assumptions about intelligence and its relation to world. This includes a dominating role for representation of concepts and logical manipulation of those representations to arrive at veridical claims about the world. Most AI is generated with logic or constraints focused on “correctness” of behaviors and inferences. However, if emergent concepts are heuristics, and

¹²¹ Allostasis has been proposed as another driver (Sterling 2012). The homeostat was incapable of the anticipatory requirements of allostasis, but the concept warrants an emphasis on performative intelligence every bit as much as homeostasis.

¹²² *Research: Consider how different prediction methods may be analogs to performative systems and methods, which could enable more capable systems. How does homeostasis (and allostasis (Sterling 2012)) factor into current AI/ML approaches? Can it be measured for all models as a complementary objective to maximize over different time scales?*

¹²³ This is also known colloquially as “GOF AI” (for “Good Old-Fashioned AI”).

those heuristics (or the constructive process thereof) persisted across our lineage due to effectiveness in coping with the world, then the relationship of intelligence to truth is at best indirect (Hoffman 2019). Indirect relationships should not be avoided (truth is often very effective), but the evolutionary process that shaped us suggests that there are key elements under-served or missing altogether in AI initiatives. Hubert Dreyfus (Dreyfus 1972) was perhaps the first and best-known critic of Symbolic AI. While some of the initial criticisms have proved incorrect, the core concerns remain valid and unaddressed. Again, there are potential exceptions, including advances in Reinforcement Learning, which I will come back to.¹²⁴ This chapter describes the importance of interaction and performance, which are core to agents adapting additional capabilities, including for symbol manipulation and reasoning over representations of the environment.

For humans and other living systems, **perception is not fundamentally driven by getting it “right”; it is shaped by fitness, including the empirical frequency and impact of the perceptions that evolution and learning create.** Dale Purves and Beau Lotto provide compelling scientific evidence for this case, which they call the Empirical Theory of Perception (Purves and Lotto 2003). Likewise, Donald Hoffman provides both research and models demonstrating his Fitness Beats Truth Theory (Hoffman 2019).

Where Dreyfus emphasizes the implicit performative character of living, Purves and Lotto explore evolutionary processes based on cognitive research. The results present a clear picture; perceptions are heavily shaped by what enabled propagating genes given their environment (i.e., what works) over accuracy.^{125,126,127} Purves and Lotto use optical illusions as tangible examples of how our perceptions are shaped by their influence on our coping in environments we evolved in.

¹²⁴ As discussed in a later chapter, I consider Reinforcement Learning and Evolutionary Methods as key capabilities along the path to Artificial General Intelligence (AGI). However, I believe these methods still make assumptions that are too narrow to achieve the performative requirements of naturally intelligent systems.

¹²⁵ *Research: More speculatively, consider if the concepts of veridicality, accuracy, and precision would lose any fundamental meaning if we were able to trace cognitive adaptations backwards. Can one make sense of them as derived from something more basic, such as consistency? If so, it could be that consistency first becomes valuable as an adaptation derived from “what works”, and only from there do concepts like accuracy become valuable.*

¹²⁶ *Research: Is the ability to generalize consistent with the Empirical Theory of Perception? What about the ability to simulate events? Consider learning to break when driving over curved roads. There is no need for empirical evidence in the form of crashing to learn how to “perceive” these adaptive behaviors (too recent for direct evolutionary adaptations). Is explicit instruction required to learn to break without first crashing or is our implicit sense of the effects of acceleration and centripetal force necessary and sufficient? While driving a car, and the associated high velocities, are new, it does seem that we are applying some adaptations to learn to implicitly control them. We also have the general ability to self-correct with new information. It is also worth keeping in mind that what we invent is constrained by our past capabilities (i.e., we might not have built cars if we did not have the built-in potential to sense and learn to avoid crashes).*

¹²⁷ *Research: Karl Friston claims that what distinguishes life (and some nonliving processes) is the process of using inference to minimize surprise (Friston 2010). Consider if/how such a process might guide the Empirical Theory of Perception. This would create attractor states for inferences given observations (one of Friston’s claims) but where this is driven by empirical frequency of encountered phenomena and shaped by evolutionary fitness. This requires creating an environment for the systems and having the population adaptations guided by a fitness function. If the environment is similar enough to the ones our ancestors evolved their attractors in, then we should see similar inferences (e.g., the same color-based optical illusions).*

They are clear byproducts of evolutionarily adaptations that make no sense outside the context of evolutionary fitness constrained by the particulars of our environments. Evolution-driven heuristics literally and figuratively color our perception of the world.

Even though the world we each construct may only make sense with respect to evolutionary history, our perceptions should largely align due to similar evolutionary constraints for our ancestors. This suggests we can better explain our perceptual and conceptual constructs from an evolutionary need.¹²⁸

In more patent cases, such as the examples provided by Beau and Lotto, we can discover parsimonious explanations for the evolutionary history that led to particular adaptations.¹²⁹ Continued identification of explanations is incredibly valuable research. However, evolution, as an emergent phenomenon, is an irreversible and incredibly complex process unconstrained by any need to describe itself (let alone in terms we understand). I doubt we will be able to reverse engineer the process, not only due to lack of data but because explicit decisions break down, as discussed in Chapter 1.

What does this all have to do with AI? **The ability to adapt (over a range of time scales) is as critical to intelligence as is prediction and inference. Indeed, prediction and inference likely evolved in service of adaptation.** This is not only underemphasized in Symbolic AI. Consider Supervised Machine Learning (ML). **Rather than implementing a method for machines to adapt behaviors that work (i.e., cope with the world), Symbolic and Supervised AI are built and trained, respectively, given human-derived correct answers. The derivation of these correct answers required the effectiveness-driven adaptations of the human modelers to evaluate and drive the machine's adaptations. However, it is very superficial products (rules or labels) of these processes that are input to the AI models. Not only is this indirect (minimal interaction involved in receiving labels for learning), but it skips most of the learning process that derived those results in humans. For AI to acquire sufficient information to develop human-like perceptions and cognitive processes requires rich interactions with an environment with requisite complexity. Would we recognize what is in images without years (or millennia) of interacting in and adapting with the world? What grounding emerges from our adaptation process compared to learning in this relatively superficial manner for AI/ML? Moreover, images are a cheap substitute for what natural systems learned from to develop vision. It seems unlikely that sufficient information is available in a set of labeled data for AI/ML, no matter how large, to perceive beyond superficial and brittle classifications.**

Here, points made in the Chapter 1 become important in our future AI research program. The outputs from machines are classified by us. They are therefore emergent to us, not necessarily the machines. Extraordinarily little of our grounding could be conveyed to the machines through

¹²⁸ Research: If there are statistical or logical patterns with respect to empirically driven theories in evolutionary adaptation events, then consider if/how we can start grouping percepts and concepts based on them and possibly even mathematize them as such with respect to Shannon, Semantic, and Effective Processes (discussed later).

¹²⁹ Research: Gestalt research is a rich source of emergent properties with conceivable evolutionary advantages. Determine a means to develop and assess AI agent adaptation of Gestalt effects. How would this be measured from agent internal states?

labels to learn how to process images. AI/ML models are often trained with data features that *represent* emergent concepts but not with any of the implicit properties projected onto the data from the mind of the observer who defines the data (e.g., a binary value representing a victory or influenza, etc.). In Symbolic AI, the assumption is that underlying properties that evoke the concepts are unnecessary. Only the symbols must be modeled, as well as processes by which to manipulate them directly. For Supervised ML, it is expected that the training process that is shaped by the concepts (as labels) will learn to evoke them with underlying processes analogous to those used by humans to evoke and map the concepts labeled in the training data. However, the means by which representations are provided bear little resemblance to how the analogous features are provided to living systems. While humans do learn through explicit correction, this is not the only process, and is rarely learned passively when used. For example, humans often manipulate objects to learn how to ground them.

Note that the issue of concern here is not that Supervised ML cannot perform the types of reasoning that humans can given images. Even if higher reasoning were made possible, say through new supervised ML architectures or coupling them within other reasoning methods, it would not alone address the need to interact with the objects/events they are meant to reason about. Not only that, but the interactions (both the types and in some cases the propensity of certain sequences of interactions) evoked with the objects must be driven by past effectiveness and have the potential to influence effectiveness of future actions.¹³⁰ Reasoning must be grounded, and the grounding properties evolved/learned are strongly influenced by their effectiveness over evolutionary and lifetime history.¹³¹ I therefore group typical Supervised ML along with Symbolic AI into ‘Representational’¹³² methods, or methods that emphasize the modeling/learning of representations based on how veridical they are to an outside modeler’s own grounded inferences.

How symbols in the world become grounded depends on how they are used. **It would be highly unlikely that Supervised ML, such as training classifiers using labeled images, learn how classified objects are grounded for humans. How humans ground objects, and perceptions more generally, depends on their role as affordances (Gibson 1979). The role of affordances is too limited for ML that passively observes labeled images to learn how to classify objects from them. The role of entities as affordances underlies our Reasoning processes as much as Sensory processes.** Our reasoning capabilities likely evolved as extensions to lower-level sensory processes. To the extent that we experience concepts as we reason about them, it is self-evident

¹³⁰ An example of the latter might be an action by humans to migrate north, which could create new evolutionary pressures that change the effectiveness of future actions, such as cultivating certain types of plants. A more relevant example for this chapter might be getting stung by a bee, which influences one’s perception of honey as a food source, which might in turn increase the effectiveness of actions to build certain types of tools.

¹³¹ This is likely not only true for how individual entities are grounded but also how effectiveness shaped a more general grounding “space” that in turn shapes how grounding relates across entities.

¹³² However, not all supervised methods fall under what I later call “Explicit Representational Methods”. I will argue that some methods at least have the capacity to ground symbols, whether supervised or not. The concern for this chapter is the quality of the grounding given how such models are trained.

that grounding is involved in these “higher levels” of intelligence. That grounding must be shaped by how effective it is for coping with the world.

This challenges our ability to simply engineer in the heuristics humans apply for inference and reasoning. It seems likely that theoretical explanations for heuristics require not only an explanation of the mental properties but also the evolutionary history for how that emergent property was adapted. This means it may be insufficient to simply capture grounding through interacts with entities that we want AI to reason about. A model may need to account for the implicit features in the evolutionary history that shape the emergent properties evoked by the given entities of interest.¹³³ All the labelled data in the world would not alone train AI/ML to learn to perceive objects like how humans perceive them.^{134,135,136,137}

Evolutionary dynamics are not the only limitation in learning purely from images. **As concepts become increasingly abstract, context is increasingly important, and the use of labeled data available becomes too limited. What is evoked in humans for a given image is influenced by much more than simple labels or captions, and it varies by a much greater breadth of context than what is available in sets of individual images. While humans can classify with images, what they perceive from them is shaped by much richer interactions with the real world. Thus, there are dynamics not captured by images or labels, including events where observers are actively involved in scenes as they unfold.**

Another issue, particularly for Deep Learning, is that its large degrees of freedom might be used to learn non-sensible models. Deep Learning may often be more similar Ptolemy’s use of epicycles to model planetary movements than the Newtonian model of planetary orbits. That is, some trained models may not contain any real insights about the world they observe and

¹³³ It may also need to interact with many different entity types, since we do not know how grounding is shaped across entities.

¹³⁴ *Research: Determine (theoretically or empirically) if it is really the case that evolutionary constraints require more than human labeled images for supervised ML. Is this possible to prove/disprove? In what ways can evolutionary history be made irrelevant through the indirect constraints placed on the labels that humans provide? What specifically cannot be captured this way? What evolutionary data would help? What are the minimal interactions required? Empirical results demonstrating the need for evolutionary information would change the general research trend and its implicit plan for advancing computer vision and other AI/ML areas.*

¹³⁵ One could argue that assuming AI should ground concepts like humans limits how (and possibly how far) AI advances. I agree that there is no reason to *only* build AI constrained by biological intelligence. However, AGI will require grounding similar to humans due to the all-encompassing nature of our emergent worlds. Certain forms of (useful) superintelligence may also need to first be capable of AGI.

¹³⁶ *Research: Are there any implications given the Empirical Theory of Perception for the relative frequency of object/event classes sampled for AI? Would different object/event sampling biases (e.g., many cat videos vs. the number of cats humans tend to come across in real life as a proportion of time) impact the “strategy” learned by AI/ML to predict different classes? Consider comparing generated images from the same models trained with different object/event class imbalances.*

¹³⁷ *Research: Determine a research project that aims to extract as much information as possible for what goes into a human’s labeling of more abstract concepts given data (or any labels that currently challenge AI/ML). Also, consider what a child would need to understand such concepts. Are there concepts that must first be learned prior to learning others, as suggested by the Zone of Proximal Development (Zone of proximal development n.d.)? How can AI/ML Curriculum Learning methods (Hacohen and Weinshall 2019) be improved based on both Psychology and Evolutionary research and theories?*

predict; they just contain enough free parameters to often achieve similar results. As universal function approximators, they may excel in cases where there are many different functions that could approximate some “real” one.^{138,139} Interaction provides an opportunity to experiment with objects/events, which constrains models.¹⁴⁰

Of course, computer vision has made incredible advances in recent years. Methods that improve over the state-of-the-art continue at a rapid pace, and it would be foolish to suggest that they will not continue to overcome some proposed limitations. Moreover, alternatives offer nothing drastically new. However, it is also good to consider what assumptions about Intelligence might be invalid or incomplete to further accelerate progress.

Emerging Veridicality

Supervised learning assumes a truth about what images represent as expressed in a given loss function. I do not want to make any strong claims about the ontological status of Truth, particularly since I do not have any strong ideas for what Truth could emerge from. Math/Logic are perhaps the most fundamental concepts, making them exceptionally difficult to talk around to determine implicit properties that might account for them. However, if there is some Platonic form of Truth “out there”, then, as evolved systems, it is frankly puzzling how we came to know it. There is therefore value in attempting to have AI emerge the concepts themselves without any predetermined constraints that assume use of Logic or Math at the level of the features that make up the AI’s environment. This hypothesis relies on the assumption that the environment contains enough regularities that truths conceived will tend to converge within and across agents, just as they have for humans. I fully understand that it would seem incredibly inefficient to force AI to learn Logic, when we already have essentially complete methodologies in place that can be implemented. However, just as optical illusions have shown with humans, it is not absurd to hypothesize that the emergent concepts constructed by AI and the contexts by which they are mapped to observations must be learned. What is conceived as “red” should depend on perceptions shaped by effectiveness. **Logic/Math may be universal. Consider that, for the illusions we know of, inductive and deductive logic enabled us to determine that they are illusions. However, we do not know the impact of starting with Logic/Math predetermined for AI. How might this affect how it grounds and otherwise copes with its environment, including**

¹³⁸ Scare quotes are used here for “real” due to the complications of assuming what is “out there” to be modeled can be described as a function. It might be more appropriate to place the quotation marks on “function”, but this then complicates what it means to discuss universal function approximation. Given these thorny issues, the point made should be understood without claiming any formal assumptions about models or the real world.

¹³⁹ *Research: Is there any reason to believe that our brains are also primarily exploiting their large parameter space for prediction and inference? If so, then how do we come to understand and provide parsimonious explanations for the systems we conceptualize? Understanding the difference between how our neurons model systems and how we explain those systems seems important for building Explainable AI, as well as understanding what it means for us, and AGI, to model our world.*

¹⁴⁰ The experiments that tend to be performed may also need to be biased towards the types that humans apply to cope with their environments. For example, the risk tolerance for different contexts, what different objects tend to be manipulated together, etc. may be constrained, which in turn might bias the types of models that humans form for processes.

the very concepts/symbols the AI emerges? Order of intellectual capacities/biases may matter just as much as order of genetic mutations matter for what lifeforms evolve.^{141,142}

If Veridicality is a heuristic, then it is an important one, especially for humans. It enables reasoning and much of what we can communicate across a wide range of abstractions. If, as suggested by many AI researchers, humans (and other animals) do not primarily learn through supervision, then we need to explain how we have formed concepts and processes associated with conceptualizing and learning true and false, and how it is applied to our models of systems and environments. It may be that AI agents must evolve/adapt mechanisms to supervise themselves by hypothesizing and then collecting evidence from the world and/or learning from each other. That is, **AI must form concepts of true and false that improve their ability to effectively cope with the world, and they also need to evolve/adapt means to learn how to apply true or false assertions. They need to somehow emerge their own supervised processes, such that their explicit reasoning processes are effectively grounded.**

While this section has singled out Supervised ML, it is equally likely that Unsupervised ML requires evolutionary context to constrain how ML structures data. What regularities in data are selectively attended to, and in what ways, depends on the effectiveness of the options evolved/learned. For example, how humans assess similarity of perceived phenomena will affect how those phenomena are grouped.¹⁴³ This is not only important for comparing like and unlike objects/events, but even determining what and how such emergent objects/events are constructed.¹⁴⁴

Amalgamated AI

Research and theories on the concepts of Embodied, Enacted, Embedded, and Extended Cognition share the idea that biological intelligence involves more than just the brain and passive representations of the environment. For example, Extended Cognition suggests that an agent's cognition extends out into the world beyond the brain. Mark Rowlands (Rowlands 2010), makes compelling philosophical claims for Embedded and Extended Cognition, which he combines into the concept of Amalgamated AI. As discussed later, this strongly suggests that **AI should be amalgamated to advance towards AGI.** Of course, the lack of interaction with the environment in many AI methods currently limits the ability for the AI to learn/evolve amalgamated cognitive processes. This can be related back to the theories on Affordances. For example, JJ Gibson's

¹⁴¹ *Research: The "Unreasonable Effectiveness of Mathematics" (Wigner 1960) also applies to Logic. "Effectiveness" is used in a different sense that I use the term. Consider how this either contradicts a heuristic claim about Truth, or how this effectiveness can otherwise be explained.*

¹⁴² *Research: How would Veridicality being an evolved heuristic relate to my claims in Chapter 1 that Concepts necessarily have truth values? Are heuristics ever necessary? Necessity would seem to assume Logic. If both claims hold, then what does this say about the consistency of Concepts? Can they be both precise and contain unresolvable paradoxes at their level?*

¹⁴³ This relates to Gestalt principles (Todorovic 2008), which are discussed in Chapter 4.

¹⁴⁴ *Research: Is there a more effective way to cluster data without either limiting them to our existing categories and concepts or passively learning them? Determine a model and environment requiring agents whose perceptual process can be optimized depending on how they influence the effectiveness of behaviors, such as avoiding dangerous areas or collecting certain objects deemed to be of the same type.*

Ecological Theory of Vision (Gibson 1979) posits that the properties of percepts are determined through actions of the given agent on the environment (e.g., moving in space) to make information available (e.g., in the optic array). This Ecological Theory contrasts significantly with the purely representational (and possibly even a computational) focus of AI research. As Mark Rowlands (Rowlands 2010) suggests, these processes are not necessarily exclusive to internal transformation rules happening inside the brain (e.g., in various visual brain areas).

Again, most of data we are using for training AI/ML simply does not capture the types of information available to living systems. Humans leverage their ability interact with the world to seek out more variables to distinguish possibilities. They mentally and physically explore to find explanations for conflicting beliefs or ambiguities in perceptions (at least when doing so is itself an affordance for the given agent). From saccades to object manipulation, it is their performance with the world that exposes information. In fact, humans may have heuristics to cope with uncertainty beyond simply gathering more evidence. They may have particular ways to explore which are highly effective for the types of environmental variables most relevant during their evolutionary history. Interaction with an environment is also the most powerful method to infer causality. If this is critical to our own perceptions, then its general absence in how we train computer algorithms, where only passive and internal transformations are learned, will limit the ways in which they advance and can be utilized.^{145,146,147,148}

Although I have mostly discussed heuristics in terms of emergent concepts, **cognitive processes are themselves heuristic adaptations. For example, how we learn, what we remember, and how we solve problems, have all been adapted. Adaptations for how we interact with the environment could be considered meta-performative adaptations that are also performative. This is an amalgamated process that optimizes how we amalgamate environment dynamics and cognitive processes. Heuristic processes adapted over evolutionary history matter for how we learn with relatively little data throughout life, including for learning new concepts.** For example, the ability for humans to learn to classify and comprehend unfamiliar and rare stimuli is integral to how we cope in the world, as well as how we modify and reuse existing concepts (e.g., there are an infinite number of sentences that can be communicated, and we understand many combinations we have never heard). As discussed later, embedding methods hold promise in this area by learning an approximate metric space (known

¹⁴⁵ Research: Consider if there is a scalable way to enrich the imagery or video we already have available with the types of phenomena Gibson's proposes humans leverage for perception. This would enable leveraging a source of potential information that is currently not used by state-of-the-art algorithms that train on passive images or video.

¹⁴⁶ Research: Determine requirements for RL or other agents to meet the criteria for "amalgamated AI".

¹⁴⁷ Research: A way to bring together Heideggerian (Dreyfus 2007) and Bayesian AI is to focus on developing bio-inspired coping with Uncertainty. Determine how to best tackle problems for which our priors are minimally informed. Do we simply try to optimize our understanding, or do we have heuristics that help us perform despite (or by even leveraging) uncertainty?

¹⁴⁸ Research: Consider developing a taxonomy of adaptive behaviors based on empirical evidence, and then use it as a performance metric for AI/ML agents. These agents are optimizing capabilities rather than some overarching objective, which may be impossible to find, vary widely, and lead to a fragile solution. It is possible that any complex system involves multiple optimization problems (if optimization can even be defined effectively given each complex system's behaviors).

as a Semantic Vector Space) used for analogical reasoning. For this section, I want to point out that the value of interaction is still untapped in training these methods (there are partial exceptions, such as learning latent¹⁴⁹ spaces via Deep Reinforcement Learning) and hold as much if not more potential for building effective Semantic Vector Spaces.¹⁵⁰

Ethological AI

The above sections emphasize that all emergent percepts and objects derived meaning through adaptive, evolutionary performative processes. Since humans developed heuristics for effectively grounding stimuli, a question is whether AGI must go through similar evolutionary pressures. A challenge is that many contingencies can influence the particular grounding that evolves for agents over many lineages interacting with particular environments. Moreover, emergent concepts presumably build off each other, creating a messy implicit evolutionary mapping "schema". How can we guide such processes to not only be effective but to do so in a way that grounding for AI agents resembles what humans evolved and learned? In what circumstances is this necessary? In what cases might this be deleterious?

What sort of virtual environment might be sufficient to enable grounding in AI that is sufficiently close to human grounding to be useful for future applications (including scientific insights into our own minds)? How close do the environments need to be to our own? How sensitive are the evolutionary paths to historical contingencies?^{151,152} I suspect that answering these questions is a long-term research effort. Regardless, providing a more fundamental relationship with the environment from which information is processed is a good start. For example, it may facilitate AI learning more abstract concepts. How much specific percepts and concepts differ might be manageably controlled. One could call this general research approach, "Ethological AI". An example outcome would be learning similar optical illusions due to some driving force that requires forming heuristic assumptions about, e.g., color constancy in different brightness settings. The field of Ethology is motivated by the belief that understanding organisms requires studying them in the wild. This is not only important for sociological studies, but also psychology and neuroscience (see for example, (Platt Labs n.d.)). If the study of many animal behaviors requires that they be observed in natural conditions, then one would expect that

¹⁴⁹ "latent" here is used synonymously with "semantic vector" or "embedding", which I discuss at length in later chapters.

¹⁵⁰ *Research: For many problems, no reasonable amount of sampling will suffice in training, given the curse of dimensionality. Consider how generating/improving lossy emergent heuristics relates to the tradeoff between data requirements and approximations, including how we encapsulate entities into subsystems by treating them as well-behaved emergent concepts. Also consider how analogy-making is leveraged. In AI, agents must be able to group phenomena into patterns through continuous analogy-making. Consider embodying different sampling methods in agents that interact with and co-evolve in simulated environments.*

¹⁵¹ Although it is beyond the scope of this manuscript, there may be ways to relax how close the environment for AI must reflect the environment our ancestors evolved in, although to what degree is an open question.

¹⁵² Another fundamental issue is that evolutionary modeling methods require explicit predefined fitness functions to guide selection processes. This seems inappropriate for modeling biological intelligence and may be insufficient for building AGI for the same reasons.

organism adaption and learning also depends heavily on environmental conditions. Therefore, AI may also need to adapt/learn “in the wild” to achieve human-like capabilities.¹⁵³

Evolutionary Algorithms and Reinforcement Learning could in theory provide requisite interactions, but research must involve addressing how to feasibly place similar pressures on agents as provided in our own evolutionary history such that similar heuristics emerge. Admittedly, this feels like solving AGI the hard way. However, if the needs are recognized and promising paths forward for true AGI are proposed, then I am optimistic that resolutions for practical issues such as these will follow, just as they have for other AI methods (e.g., ANN backpropagation) when the needs were too strong to ignore.¹⁵⁴

Ultimately, an imperfect but flexible solution may be valuable. Even if the evolutionary pressures differ, Ethological AI research would provide a history from which one can explore how agent perceptions are influenced by their effectiveness. Moreover, some types of emergent concepts may end up being robust. For example, emergence of veridicality and causal inference may be generally effective across conditions, with differences being informative about our own fundamental biases and selective pressures.^{155,156} Agents do not construct concepts arbitrarily. Reinforcement Learning should converge on reliable concepts through trial-and-error to be effective because the ranges for variables is limited by the particulars of the world, and the precision required to survive could be considerably less than perfect. One can effectively get to

¹⁵³ *Research: This includes the effects of providing an information-sparse environment. For example, learning to perceive a tiger attempting to hide could be implemented in an agent that comes across tigers in a simulation, and must learn to detect them, including with various degrees of occlusion, lighting, etc., to avoid them. The agent may need to have realistic motor action options to make the effect of perceiving the tiger or not even more aligned with how it developed for different prey.*

¹⁵⁴ *Research: Consider how to develop multi-scale, multi-environment iterative processes.*

¹⁵⁵ *Research: Develop an Evolutionary Algorithm or Reinforcement Learning problem for which supervised learning emerges for the agent to cope with the environment. Determine some way to measure agent classifications. Control environments to determine if this changes classifications for the same objects. For example, known optical illusions could be used to see if the agent always perceives the same color for the same object. Then, use these agents to label data to train a supervised classifier and assess how close the classifier is with having the historical experience compressed to labels. What types of labels converge across environments? What else could be provided that would improve performance? Class frequency as a parameter may help measure how percepts change with importance.*

¹⁵⁶ *Research: With a temporal dimension, the role of performative cognition becomes that much more important relative to representational cognition. Consider using something like Deep Recurrent Neural Networks (with Attention) but perhaps where the impacts of previous timeframes are weighted by (Shannon) Surprisal.*

something worth calling “truth” naively. We may be in the ballpark if there are bounds to variables (“in the ballpark” meaning sufficiently effective).^{157,158,159,160}

Conclusion

All the issues raised in this section point to the need for AI/ML to interact with the environment, not only to construct effective perceptions but to learn to effectively reason. Humans solve problems iteratively with the environment. They use some implicit level of confidence for asking questions or taking other steps and receiving feedback for how this informs next actions. Planning crosses many timescales based on implicit or explicit objectives and past experience, including with other intelligent agents. AI/ML agents must have a more fundamental

¹⁵⁷ *Research: If the research above enables agents to learn representations guided by fitness, including recognizable optical illusions, then there are various research concepts to consider incorporating, either as enablers or guideposts for what agents adapt. Examples include determining how to develop an evolutionary account of Dreyfus’ claims (Dreyfus 1972), Freeman’s Neurodynamics theory (Freeman 1999), and how to apply Gibson’s ‘affordances’ (Gibson 1979). For example, Dreyfus’ as-structure concept (Dreyfus 2007) would likely need to emerge from evolutionary coupled-coping. Also, determine how Freeman’s Neurodynamics theory applies for assessing agent internal states as processes, or using concepts from Force Dynamics (Talmy 2000) to guide the relationship between actions, environment, and internal states. Freeman critiqued the representational assumptions of Symbolic AI, so his research may provide insights for how natural brains learn to cope with the world. Also, develop an evolutionary or RL environment that enables learning new ways to use affordances (e.g., maximize number of ways to use a hammer in the environment), and ways to combine objects in the environment to increase agent effectiveness.*

¹⁵⁸ *Research: Rather than evolving directly for representation, brains may have evolved for agents to rapidly try out mental “phenotypes” on the fly or over time to adapt within a lifespan (rather than across lives). Neural dynamics as gene proxies would manifest a large set of behavioral repertoires as phenotypes. Parallelism of neural processes may provide a means to “simulate” parallel phenotypic behaviors (“simulation” is not the right word here, since the behaviors are actualized. Instead, it is a set of gene proxies that can be easily adapted during life rather than requiring new lives to sample genetic sequences). Genes do not manifest truths about the world, so this concept simply extends gene functions for exploring phenotype effects on the fitness landscape to neural processes for exploring phenotype sensorimotor effects on the fitness landscape dynamically. Consider how to test this. For example, is there a fitness or information gain curve that maps from agent gene-driven to neural-driven phenotypes? Neural processes are themselves phenotypes of genes, so the distinction may be difficult to separate. An alternative way to conceive of the relationship could be as a hierarchical dependency (i.e., gene-driven phenotypes enable neural-driven phenotypes).*

¹⁵⁹ *Research: An AGI algorithm may require Reinforcement Learning (RL) to make decisions and learn value but where each RL agent goes through natural selection to mutate to the greatest rewards for given states, including evolving functions to choose the reward functions (i.e., a meta function to optimize learning), which enables evaluating things never seen before via analogies to known state/feature properties, etc., given current context. This leverages the power of RL models (e.g., Bellman’s equation, Temporal Difference learning, Deep Learning approximation variants) while still being able to adapt, and in a generalizable way that adapts to the context of a given problem. Parameters, such as discount factors should also vary given context, so a meta function is required to determine how to determine parameter values to use for a given problem (i.e., context). It makes sense to let an RL agent apply different parameter values, such that it can vary by state when evaluating a policy.*

¹⁶⁰ *Research: Just as humans are used as the gold standard for representational AI (e.g., Expert Systems, labeling images for Classifiers), consider how they could be used as the gold standard for performative AI (e.g., adaption), particularly when applied to complex systems. What heuristics might be transferred, and modified to scale, to allow that same level of short-term predictability and adaptability? Is there anything here that can extend beyond Learning from Demonstration methods (Schaal 1997) to enable AI/ML to cope in more abstract environments? If so, this may lead to human-computer relationships that take human heuristics and apply them (or even discover ineffective ones) to different areas to solve novel problems and learn to cope with novelty.*

relationship with the environment in which the objects observed are actually present and respond to agent actions to learn about them and the affordances they may offer. This requirement is theoretically met by Reinforcement Learning and Evolutionary Algorithms.¹⁶¹ However, having requisite complexity in the types of interactions and environmental pressures sufficiently similar to what humans dealt with, so adaptations resemble the types necessary for AGI, must be considered. Interaction is a critical feedback loop for nature (across generations for evolving and across a lifetime for learning). We need to focus efforts on transforming data into interactive and proportionally impactful environments for the agents, including with other agents.^{162,163}

Reinforcement Learning has become incredibly powerful since scalability was improved through Deep Learning (among other enhancements). We also need breakthroughs for simulations within which to train them.¹⁶⁴ One challenge, however, relates back to the Description Gap described in Chapter 1: Measures for effectiveness are not easy to make explicit because effectiveness ultimately plays out at the mechanistic level, where functional concepts do not apply. Although Darwinism may provide an effective emergence heuristic, ultimately grounding may need to be tied to maintenance of mechanistic structure and dynamics over time.

Another potentially important consideration for adaptation is how we have evolved to cope as far-from-equilibrium systems. Living systems thrive at the "edge of chaos" (Langton 1990, Crutchfield and Young 1990). It has been claimed that being at this edge is what enables organisms to maximize information gain (in terms of efficient computation (Langton 1990)). It may also facilitate context switching, which may resemble more broadly effective coping with

¹⁶¹ However, it is difficult to see how such an immense space of environment states and actions could be defined as a grand RL policy. The size of such a policy may be too large to even effectively approximate. If so, then this provides even more reason to better understand and apply performative intelligence like biological agents.

¹⁶² *Research: Consider how to replace observation-only image and video data with models of the same data but capable of eliciting positive and negative rewards and alternate histories that depend on virtual interactions with the ANN agents.*

¹⁶³ *Research: Consider an extension to Active Learning (Settles 2010), where instead of inferring once per sample (e.g., image, text), the algorithm samples multiple times after blocking out parts of the sample to present the modified samples to humans. If the blocked parts lead different humans to apply different labels (or at least no longer the original label applied), then the algorithm can isolate what is essential in the image for the given class. This creates some limited interaction with the environment. Could this be used to enhance neural attention mechanisms (Vaswani, et al. 2017)? How could the results be used in a generalizable way to feed or initialize other neural networks to accelerate or improve their learning?*

¹⁶⁴ *Research: Consider other ways to use video games or simulations to provide interactive environments for computer vision learning, especially if they model physical laws, such as gravity, momentum, and collision-related effects. Agents could reproduce, allowing for accumulation of "genetic" mutations that provide mechanism-level instruction code e.g., for ANN properties. If there is more than one agent in the environment, they should be able to cooperate/compete. Fitness must be defined, such that it theoretically can lead to the survival of the beneficial mutations given environment coping they enable. The fitness should then be associated with the goals of the human developers' requirements (e.g., correctly categorize cars, but in such a way that the behavior with the cars is what is in agreement with the goals of the human trainer; not necessarily just a perceived matching of the human's labels for objects). This allows more useful relationships that account for interaction and temporal dimensions (e.g., not provided in imagery alone, and alternative histories for video data). If neural embedding methods are used, symbols that emerge from it may evolve based on what behaviors they evoke in other agents (and the agent itself).*

state transitions; something current AI does do not particularly well, except in narrow spaces with dense sampling coverage. More generally, there is a need for broader consideration for how Complex Systems insights can be leveraged to advance AI. Although there is a rich history between these two fields, there is neither a broad understanding of Complex Systems principles and research in the general AI community or a lot of research focused on implementing the many specific insights discovered about complex systems. These insights may be just as important and valuable as the insights applied from other areas (e.g., Reinforcement Learning from Psychology). While correcting the ignorance of current AI's cybernetic heritage could lead to rediscovery of important research ideas for AI, I suspect that **common knowledge in Complex Systems research in the AI community will be necessary to truly advance AI to anything resembling the intelligence of living systems.**¹⁶⁵

Finally, although I have focused on ML above, all these performative claims equally apply to logic- and explicit model-based AI (e.g., Ontologies, and Probabilistic Inference). In these models, the modeling assumption is that humans can ground symbols for AI by linking them to other explicitly represented symbols.¹⁶⁶ This is even worse than learning from labels, which can at least capture implicit features when the model is sufficiently flexible and complex (e.g., Deep Neural Networks). The next two chapters discuss these and other issues for such AI methods. The overall claim will be that the mind is not built on logic or explicit symbols. As proposed above for Veridicality, AGI must develop the ability to use explicit symbols from a more fundamental implicit base shaped by effectiveness.

¹⁶⁵ *Research: Complex Systems discoveries should be explored in the future to integrate Machine Learning with attempts to model agents at the “edge of chaos”. Consider how equilibrium states could be measured for different types of AI, even those not performed in simulations. What is required to at least determine in what ways they are or are not deficient, and what would be required to make them adapt in such conditions?*

¹⁶⁶ This statement assumes the AI model developers and users believe the models can ground the symbols modeled or that grounding can be bypassed altogether. There are also models that developers and users may acknowledge cannot ground the symbols, but that grounding is not needed for the limited set of specific model application(s).

Chapter 3: Explicit Representational Modeling

Introduction

Explicitly mapping relationships between emergent concepts is important in the Sciences, including to organize our understanding of various domains of reality. In AI, such maps are often used as models to perform prediction and inference. Many models treat these concepts as primitives, whereby variables are explicitly labeled to represent objects and relationships. This is common practice in the field of Knowledge Representation and Reasoning (KRR) (Knowledge_representation_and_reasoning n.d.). While the concerns with this approach to modeling have been discussed in earlier chapters, this chapter aims to focus more on the implications for AI rather than the philosophical perspective. Understanding the issues with these methods is important because considerable investment is going into them without a clear understanding of the importance of real grounding in models, or at least what is required for grounding. As Deep Learning methods face difficulty in advancing up the AGI chain, history is at risk of repeating itself, reverting back to these explicit models without a better understanding of grounding/emergence. We must be very clear about what did not work and why.

Moving away from Intrinsic Emergence

As a reminder, I am avoiding claims about intrinsic experience when discussing AI/ML and grounding. I explained in Chapter 1 why it seems impossible, from an extrinsic reference frame, to prove any digital model, and likely any external system whatsoever, experiences anything. To see this, we can turn to the Chinese Room Argument (Searle 1980). The Chinese Room Argument is a thought experiment demonstrating that a computer does not need to understand anything about messages received (i.e., inputs) to decode them and produce any conceivable outputs. The argument has us consider a scenario where a human replaces a computer to answer questions in a language the human does not understand (e.g., Chinese). The human is in a room and can receive written questions (in Chinese) from outside the room. They also have a book that provides instructions for what answers (also in Chinese) to provide given each possible question. Thus, all the human must do when they receive a question is look up the matching response in the book and output it. The human never needs to understand Chinese to perform this role. The point is that this scenario seems no different than what computers do for any tasks they are provided. The book here represents the program (e.g., algorithm), which is coded by someone outside the computer. While the development of such a book for answering any Chinese question is infeasible to build in real life, the fact that any programs we can build could be likened to this scenario suggests there is no reason to believe computers intrinsically understand through the programs provided them any more than to believe the human in the thought experiment understands Chinese. The cognitive processes that go into developing the program are external to the computer just as the cognitive processes that go into developing the book are to the Chinese room.

Measuring intrinsic emergence is a problem for all model types, not just the explicit models critiqued in this chapter. For example, there is no clear way to test assumptions that aggregating

ANN node activations, no matter how complex the relationships between nodes, somehow emerges intrinsic experience. Assessing whether AI models experience anything is fraught with metaphysical issues, which is likely why the question has been ignored by many who simply want to make progress on modeling intelligent systems. I understand that perspective. Thus, a major objective of this manuscript is to really understand what we mean by grounding and, given that understanding, whether it matters for building intelligent systems. If so, this is useful for most AI research, and critical for AGI research. Thus, the real questions I mean to address are the nature of grounding and how it matters for AI.

My claim is that, even though we cannot measure intrinsic emergence, we can define and measure Grounding. Criteria for grounding is proposed in Chapter 4. We will see there that, according to the criteria I lay out, subsymbolic models (e.g., ANNs) can in theory ground symbols, while symbolic models cannot. There, I will revisit the Chinese Room Argument to propose that there are certain processes that must be performed to ground the symbols received, which are not performed within the Chinese Room thought experiment but could in theory be performed by a computer.

Note that **Extrinsic Emergence is also not the same thing as Grounding**. A rich extrinsic emergent experience might arise from watching an animation, depending on the speed at which the sequence of static frames is presented, but it makes little sense to describe the system of ink forms and motors that flip them as a grounded model. Likewise, a sentence in a novel can evoke rich emergent experience in the reader but the book itself does not ground the sentence (as is evident from the drastically different experience that an illiterate person would have when observing the sequence of text markings). Similarly, rich emergent properties experienced when observing the Game of Life (Conway's Game of Life n.d.)¹⁶⁷ do not obviously inform on the grounding of the computer model itself. As will be described throughout this manuscript, **an AI model built from labels of explicit concepts does not ground the concepts any more than a book using those concepts. Such models evoke extrinsic emergence in observers but do not themselves ground the concepts modeled.**¹⁶⁸

Thus, the extrinsic emergence is a property of the observer that would be evoked by these AI models, which can differ depending on the observer. The observer can ground messages that the message generating system might not ground. Here, our focus is on the grounding capabilities of the AI models (i.e., how they ground messages) rather than what they evoke in others. If the AI were conscious, we could say their grounding would be the intrinsic experience it has. Grounding is something essentially absent for books and various other objects or systems. So, when modeling some AI, we must ask whether the model properties exhibit grounding. We will see

¹⁶⁷ Conway's Game of Life was described in Chapter 1.

¹⁶⁸ More obviously, the same argument can be made for data; just because features are labeled as representing some concept does not by itself ground them with the concept. Someone could argue that a book is more like a message than a message generator. If so, I would argue that Symbolic AI models are also more like messages.

that certain model types can evoke an arbitrary level of extrinsic emergence for the observer but are not very grounded if at all.¹⁶⁹

Explicit Representational Models (ERM) and the Description Gap

Throughout the rest of the manuscript, I will refer to **Explicit Representational Models (ERM)**.¹⁷⁰ By this, I mean AI models that are composed of variables meant to represent phenomena for which humans have explicit terms. For example, a model could have a variable labeled as ‘tiger’. It could also create dependencies with other variables, such as ‘jungle’, ‘feline’, ‘stripes’, ‘dangerous’, etc. This form of modeling is often associated with the field of Knowledge, Representation, and Reasoning (KRR) (Knowledge Representation and Reasoning n.d.). However, it is used in many, perhaps most, types of AI modeling. Until the Deep Learning revolution, they constituted the majority of AI models and research, and for many years were even thought of as synonymous with AI itself.

I need to make two claims before moving further. First, a consequence of Description Gap described in Chapter 1 is that it is unlikely we can develop comprehensive well-formed theories at emergent levels. That is, there is no reason to expect a Theory of Everything in Psychology, Economics, or even Neuroscience.^{171,172,173,174} There is no concrete foundation to which to map our emergent concepts that would not eliminate the concepts and associated arguments themselves. Secondly, if Physics is valid, we must assume that mechanistic properties in the world, both those underlying the brains of observers and the external processes that stimulate them, reliably evoke the emergent concepts experienced. It follows from these two claims that,

¹⁶⁹ Of course, as observers, we must measure the grounding of these models as extrinsic emergent systems, but this is not the same as saying they are equivalent or even correlated. However, it still might be interesting to analyze whether there are some relationships between the levels of grounding of systems and the extrinsic emergent properties they tend to evoke in humans. I have no reason to believe any particular relationships exist, but my claims do not forbid this as an empirical possibility.

¹⁷⁰ For simplicity, I do not include methods that discover explicit models directly from data (i.e., Model Induction), such as certain Causal Discovery models (Spirtes and Zhang 2016), in the ERM category. While Model Induction alone does not meet the subsymbolic criteria I set forth later for Grounding (e.g., forming a continuous metric space), they do not require an external modeler to predetermine what variables symbolize and their relationships in the same way as the models I do include as ERMs. Thus, only some of the limitations claimed for ERMs might apply to them.

¹⁷¹ This does not rule out locally consistent claims. It would be absurd to argue that there are no valid theories in each of the established scientific fields. However, the prevalence of such local consistency may be smaller than we recognize. I am reminded of how, even though so little of the dynamics in our universe can be described analytically, this is not always recognized, since it is this small set of dynamics against which our analytical efforts are applied (e.g., “spherical cows”).

¹⁷² *Research: What criteria tend to allow for at least locally consistent theories using emergent concepts? What determines how extensive the theories can be? What makes knowledge within a field more or less cohesive? Must the concepts in the theories have been effective in the past? Must consensus in how the concepts are grounded have been evolutionarily important for our ancestors? This seems analyzable. Discovering these criteria could potentially improve our processes for building theories across the Sciences.*

¹⁷³ Some aspects of our mapping processes may have evolved as heuristics themselves, further challenging our ability to construct maps without discovering historically relevant pressures as discussed in Chapter 2.

¹⁷⁴ However, there are a variety of integrated theories that attempt to systematize knowledge within and across disciplines. These are valuable for holistic understanding and insights, regardless of whether they provide a comprehensive unified picture. An example is the Living Systems research by James Grier Miller (J. G. Miller 1978).

to reliably model an emergent domain, it is insufficient to model that world at the level of the emergent concepts, because the consistency that exists at the mechanism level does not enforce consistency at emergent levels. This, however, is a claim about building *perfectly* reliable models of the world. A purpose of modeling our emergent concepts within AI is not necessarily to build perfect models of how the concepts are used by us. If it were, then AI would be impossible. Indeed, humans themselves must operate with inconsistent models, and there is no such thing as a single gold standard human brain.¹⁷⁵ We are seeking AI to be effective for a variety of purposes, and it is clear from research that some degree of utility can be attained with ERMs in limited domains. This manuscript is focused on the quest to build AGI. How must AGI be modeled to apply our emergent concepts sufficiently enough to cope with the world, including coordinating with humans, as we do?

I believe that these two claims above (1. there is no reason to believe a fully consistent theory exists for emergent conceptual domains; and 2. there are mechanisms from which emergent concepts can be consistently evoked) suggest that different modeling approaches will have different limitations. A major claim of this manuscript is that the ways humans use concepts is driven by the grounding they evoke and from which they emerge. This grounding requires something beyond the concepts; it requires subsymbolic processes to evoke them. More specifically, the claim is that, **unless their modeled concepts map to such subsymbolic processes effectively, systems we design cannot emerge the activities/behaviors we would demand for AGI, including their ability to predict or infer the properties of phenomena about which they would be expected to reason.** My attempt to justify this claim requires other claims that are spread out over multiple chapters. This chapter focuses primarily on understanding why ERMs are much more limited than they might appear in modeling the intelligent processes of humans or grounded AI.

A practical issue with building ERMs is the manual labor involved in building them. Another criticism leveled on ERMs of the Classical AI (i.e., GOFAI) variety is limitations for modeling belief revision and uncertainty. The issue I am emphasizing here is more fundamental, and it goes well beyond KRR models, even including many Machine Learning methods. These include certain uses of Probabilistic Inference (e.g., Bayesian Networks), Logical (including Fuzzy Logic), Game Theoretic, many Agent-Based and System Dynamics models, most Cognitive Architectures, and more.¹⁷⁶ If a model is designed with explicit concepts as the modeling primitives, then it falls under the category of an ERM.

In what follows, I use **Semantic Ontologies** (Ontology (information science) n.d.) (Ontologies for short) as the ERM archetype. Ontologies capture knowledge for part of the world (i.e., domain of discourse), some using Subject-Predicate-Object triples, where the Subject and Object can be

¹⁷⁵ Likewise, there is no exact model for how concepts are grounded and used by us, since we may ground concepts differently from each other to some degree.

¹⁷⁶ It is important to understand that the issue is not the algorithms used but the types of variables they are populated with. For example, there is nothing wrong with the way Bayesian models perform inference. The issue I am making with ERMs has to do with the use of explicit labels for emergent concepts. It just so happens that the types of models listed here are often applied directly on explicit labels, which limits grounding and performance.

thought of as nodes and the predicate as an edge that links them to each other. Nodes and Edges are labeled with explicit concepts, sometimes also including explicitly labeled attributes or metadata for those concepts. An example of such a triple would be *<dogs>*, *<chew>*, *<bones>*, where *<chew>* can be represented graphically as the label for an edge between a *<dog>* subject node and a *<bones>* object node. Depending on the other concepts and relations represented in the model, the same node can be either a subject or object. To be fair, most Ontologies are not built with AI in mind, but as a means to organize information, often to make it searchable. If the engineers' objective is to use an ontology to store knowledge, then evoking semantics in (human) users may be the only intention. However, as Ontologies can be used as schemas for inference, such as for Knowledge Graphs, which are commonly used for AI, they do epitomize a philosophy in AI for how to represent concepts to reason over them.¹⁷⁷

ERMs Lack Grounding

Grounding has been thought of as that which mediates the selection of referents given observations (Harnad 2007). One might say that, if a message evokes a receiver to select the same referent when receiving the message that would be selected given direct observation of the referent, then the message is a symbol that is grounded for the referent. While referents can be thought of as actual things in the world, we only ever have intrinsic access to the grounding *properties*. Thus, we can only assume there are referents, and it is the relation of the grounding properties evoked given the referent that make that particular grounding functional. In other words, the properties of the grounding matter for how one copes with the aspects of the world they are implicitly assumed to reference. Grounding properties are thus functional by virtue of the referents they "select". We have no extrinsic access to Meaning as intrinsic experience, but we can assess the associated functionality as Grounding. Through this assumed functional relation of Grounding to Referents, I will focus on grounding properties, or what I will often refer to as the "grounding" evoked by stimuli/messages.

I discuss grounding criteria in the Chapter 4. This chapter discusses what grounding might mean for ERM. As ERM deal in labels, they are limited to associating those labels with other labels. By this, I do not mean that ERM are limited to the text domain, but that their means of interacting with any domain (e.g., imagery) is mediated by labels for explicit concepts. The means by which humans select referents given stimuli are clearly not limited in this way. Properties beyond the labels are associated with symbols. For example, "tiger" as an observed text could elicit a mental image or maybe even some indescribable experience associated with tigers. While the latter may be difficult to communicate, it can still help us cope with different referents. Words are not the whole of cognition, and the relation they have with it is complex (consider the feeling when a word is on the tip of your tongue). That is, our linguistic processes are not mediated only by the names of explicit concepts and explicit logic. If purely explicit observation and referent selection is ever used by humans (which I believe unlikely), then it is very much the tip of the

¹⁷⁷ There is no official requirement, at least that I know of, that Ontologies must use explicit concepts as primitives. As a broader definition, one could refer to the grounded Semantic Vector Space models discussed later as ontologies. However, the common usage refers to ERM, and this is how I use the term in this manuscript.

semantic iceberg. The question for AI, then, is what kind of data, model variables, and interactions with an environment are required for effective grounding?

Just as for the novel and animations examples above, if many ERMs feel grounded, it is because we, as observers, have grounded concepts evoked by the labels. ERMs might be argued to differ from these examples in their structural properties. For example, perhaps concepts become grounded in the other concepts they are linked to by modelers. However, if each node (and edge) is just another label,¹⁷⁸ then is there really any grounding produced from these structural constraints? If one were to take some written text and restructure it into an ontology, why would that make it any more grounded? In ERMs, the only information contained about a node is the ASCII characters of the labels and perhaps some logic-based operations formally implemented in the model code. Not even the words formed by the characters are grounded, which is another extrinsic emergent phenomenon experienced by the observer of the label. As I will explain in Chapter 4, the only way in which ERMs, including Semantic Ontologies, could be described as “semantic” is in what they are able to evoke in human observers. As an AI model, they have no intrinsic semantic properties. Any model that, like Semantic Ontologies, only models concepts explicitly will not itself possess semantic properties for which to ground the concepts. If grounding serves a function in mental processes, then the potential for such AI models will be limited accordingly.¹⁷⁹

Hubert Dreyfus, referenced in Chapter 2, also raised the above concerns (Dreyfus 1972). It took many years for mainstream AI to realize how important his insights were. From (Hubert Dreyfus's views on artificial intelligence n.d.):

“The human sense of the situation, according to Dreyfus, is based on our goals, our bodies and our culture—all of our unconscious intuitions, attitudes and knowledge about the world. This “context” or “background” (related to Heidegger's Dasein) is a form of knowledge that is not stored in our brains symbolically, but intuitively in some way. It affects what we notice and what we don't notice, what we expect and what possibilities we don't consider...”

Dreyfus applies various Heideggerian concepts (e.g., what he translates as ‘Ready-to-Hand’) that describe different ways in which objects in our world are understood by how we use them (e.g., a hammer is grounded in the act of hammering nails and constructing various objects for which I, as a hammer user, have constructed or want to construct). It is entirely unclear how an ERM, even one modeling relations, could model these implicit properties. As discussed in Chapter 2, these phenomena are too significant to ignore in how we understand the world.¹⁸⁰

¹⁷⁸ Some may have additional attributes, but these are also captured as labels and therefore do not affect the claim here.

¹⁷⁹ **Research: Consider to what degree Graph Neural Networks (Battaglia, et al. 2018) ground the explicit features represented in graph nodes and edges based on the criteria proposed in the Information chapter).**

¹⁸⁰ **Research: Should mapping methods (e.g., ontologies) focus on processes over entities? What would Process Philosophy have to offer for what such a mapping would look like? Could this improve on our current representational ontologies? What would it be?**

A counterargument may be that these claims commit the fallacy of mistaking the map for the territory (Korzybski 1933, Bateson 1972). AI is a model and abstracting directly to explicit concepts is useful in the same way that a map is useful. That is, the ERMs apply the only relevant properties for the function of intelligence, just as a map only includes the relevant territorial attributes for providing directions. I recognize that we are not necessarily attempting to model in AI all neural properties of biological agents. My claims are that an ERM model would make for poor models of intelligent agents. AI has an extra property that makes it differ from standard models; one that must be considered in this distinction between map and territory. **A model of an intelligent agent must be a model of a modeler (i.e., a map of mappers).**¹⁸¹ **This creates a specific problem for AI: Since it must be able to emerge its own maps (i.e., model its environment), we cannot simply provide it our own map (e.g., labels of concepts with structural properties that only we understand). This is what we are doing with ERMs.** As I will repeat throughout this manuscript, this builds models, but not modelers. **Without grounding, AI cannot, except in rare cases**¹⁸², **functionally perform the processes of the systems they are meant to emulate, whether it is a physical process (e.g., weather) or a mental one (e.g., deciding).**

Another counterargument might be that the grounding only matters for getting the labelled representations and their relationships correct. Thus, the AI requires humans to determine the labels and relationships but has no need of them once modeled. As will become clear in Chapter 4, I will claim that it is what the human is doing in this scenario that is the essence of Intelligence.¹⁸³ The product, here being just the explicit model, is a model only, not a modeler. Labels used in ERMs are meant to be symbols but having nothing internally to symbolize. The ontology is like a map without knowledge of a world to map it against. As a *model* for us, we can use the ERM labels and their associated structure, to pick out referents in our world because our capacity for grounding can be evoked by them. As AI, however, the ERM has no means itself to ground the labels from which it is composed. Its “world” is a closed one filled only with ungrounded labels. Without grounding, there are no processes to select referents intrinsically, and thus no actual relations to referents (and actually no world model possessed internally by the ERM). This affects the ability for the model to discover things about the world in ways we would expect. For ERMs, all things begin and end with the representations stimulated by the mind of the human user, where the only grounding occurs. Again, this can be useful to humans as an organizer of knowledge, but it is a stretch to call it intelligence, and certainly is not on a path to AGI.

¹⁸¹ Again, it is all too easy to mistakenly assume the emergent properties we project onto them as AI models are intrinsically modeled by the AI.

¹⁸² I qualify as “rare” here because there are exceptions where labels alone and the behaviors expected might be described as locally consistent. Much of progress demonstrated in the history of AI can be said to have proved this. Local consistency, which we might describe as exceptions to the rule, are insufficient for what we want with AGI. If the claims in this manuscript are correct, then they have not demonstrated direct progress on the path to AGI.

¹⁸³ The functional role grounding has in relation to explicit concepts goes in both directions. Grounding is used to build explicit models, and explicit models are commonly used to evoke grounding. Chapter 4 will elaborate on this point.

Finally, it is important to note that, although interaction with the world may be necessary for grounding, it is insufficient. Some Logical and Probabilistic ERMs (e.g., Bayesian-based models) have labelled variables with access to external stimuli. For example, an ERM that predicts stock prices alters those predictions given real-world data (e.g., current prices, interest rates). They can therefore operate in the world. However, I will claim that the ways in which they operate in the world do not meet the criteria for being grounded in the world. The rationale for this claim will have to wait until later, but one can get a gist of the claim by questioning whether such models have referents any more than non-probabilistic ERMs. My argument is that simply updating values and associated probability distributions for those values adds nothing if those labels and their structure are still determined outside of the ERM.¹⁸⁴

It is also of no help to point to how the computations are physically implemented. If orchestrated dynamics across the set of transistors used for calculations were sufficiently grounding a simulation (e.g., of a human walking), then we should be able to evaluate how similar those dynamics are to the physical entities simulated (e.g., a real human walking). It is unlikely that a human walking simulation should correlate in some way to the physical processes orchestrating the simulation. The hardware has no behavior with the attributes of a real human walking. Generally speaking, **there is no reason to assume that physical dynamics orchestrating the computational processes for AI agents would ground their simulated thinking in ways that resemble the thinking they are meant to simulate.** For example, the operation of logic gates to only activate when the input is 3 of something should not be conflated as grounded in the concept of “three”. What are concepts, such as “three”, for? I believe the label of a concept allows us, and the grounding evoked by receiving the labels, to move it around in mental space to set conditions within new circumstances. The grounding could also be argued to be the real mental outcomes of such changes. The labels depend on grounding through these manipulation processes. Understanding why must wait until Chapter 4.¹⁸⁵

¹⁸⁴ As an example, consider a Hidden Markov Model (HMM) developed to generate music. It is possible that long-term structure could be captured if we knew the right parameters. However, it is also possible that important dynamics are fundamentally missing because the features modeled by the HMM are all at a level above the grounding that evokes the concepts at that level. If so, then critical information cannot be representing the Markov transitions, no matter how much their model parameters are informed by external data. Humans may only be conscious of using the features modeled for the HMM generating music, but there may be critical underlying neural processes that ground those features that HMMs do not have access to.

¹⁸⁵ The problem of ERMs can also be considered with a famous logical paradox, called Buridan’s Ass (Buridan’s ass n.d.), and its implications for the Principle of Sufficient Reason (Melamed and Lin 2020). Buridan’s Ass describes a donkey that is equally thirsty and hungry and positioned equidistant to food and water placed in opposing directions. The paradox is that, if the donkey’s logic is to move to satiate its priorities, then it will not be able to decide from such logic alone and will starve. It has been debated as to whether this contradicts the Principle of Sufficient Reason, which states that all truths or facts must have reasons for them being truths or facts. What Buridan’s Ass is meant to demonstrate is that the donkey has no sufficient reason to choose food or water over the other. Something beyond the donkey’s logic (i.e., the reason) seems required for the donkey to make a choice. Leibniz resolution for why the Buridan’s Ass paradox does not negate the Principle of Sufficient Reason was that it could never occur in reality, since nothing is perfectly equidistant in the real world (Principle of sufficient reason n.d.). However, one can still in principle treat it as a potential contradiction. If so, the challenge would be to argue that freewill or a randomization process is used to make a decision. But freewill, or randomness alone, would have insufficient reason.

All the above suggests that simply linking nodes labeled as emergent concepts does not plausibly entail that the models such systems generate will accurately behave with respect to the real-world systems they model (i.e., generally intelligence systems). Thus, **ERMs have limited inference and predictive value due to their reliance on explicit variables.**

Why are ERMs Used?

It would be absurd to argue ERMs are useless, just as it would be to claim maps are useless. In what follows, I theorize on possible motivations for applying ERMs.

A. Illusion of grounding: Perhaps one reason for so much investment in ERMs, which I have already mentioned, is that it is natural for the observer to assume what the labels invoke in them is somehow also grounded in the models. This should be clear to us: When we derive concepts for physical things in the world, those physical things do not function by using those concept names. Why would we expect AI to do so? In all fairness, I suspect that most AI researchers, if asked, would readily recognize that the emergent processes that ERMs evoke are in the minds of the observers, not the ERMs themselves. If so, then there must be some assumed utility in engineering ERMs independent of whether the models themselves experience or understand (i.e., ground) the concepts with which they are populated and upon which they operate. However, I belabor the point because I claim that the models *must* be able to ground the concepts they operate on in order to operate on them in the way one would expect for AGI. I will not claim here that any particular researchers forget that the ERMs they construct do not actually understand the concepts. However, I know from experience how natural it is to assume they do when not reflecting on the points made in this manuscript.

B. Reasoning only requires explicit concepts: The most general assumption to justify using ERMs to model AGI is that general intelligence involves reasoning, and reasoning can be achieved by modeling all conceptual associations or rules. The history of Science and Philosophy might be assumed by some as largely a history of the successes of logical thinking.¹⁸⁶ Logic might be considered the pinnacle of intellectual achievement; one that should in theory trump any other cognitive processes. This leads to the assumption that human reasoning only involves explicitly interpretable concepts and logic. In such cases, the assumption is that grounding either does not

This is solved by following back to the mechanisms that generate the randomness. In computer programs, this is the code that triggers the pseudorandom process, which is sufficiently determined by how it chooses a number pseudo-randomly (indeed, the fact that we use a pseudo random process rather than a pure random process provides evidence for the span of sufficient reason, although not a proof). In humans, the process to trigger a random decision must either be found sufficiently in the brain, and/or be explained by the evolutionary history that adapted for how humans choose in such cases. Either way, sufficient reason is only obtained by requiring an understanding not provided at the level of explicit reason. This is therefore a (contrived but well-known) example of where one must model more than just the explicit variables to obtain sufficient reason. Although this is an unlikely case to occur, we would find many real examples where incorporating a different set of variables than what can be explicitly stated is required to appropriately model the situation as well as for the model to satisfy the Principle of Sufficient Reason. This therefore demonstrates at least the possibility of causal information missing in ERMs (even Causal ERMs). I must note that this does not address some other equal choice problem occurring at the lower level. However, it does demonstrate the limits of only considering a single level.

¹⁸⁶ I would disagree, since the assumption does not account for how insights give rise to the hypothesis and theories for which logic is applied.

matter at the level of reasoning or does not describe something above and beyond the manipulation of explicitly represented concepts based on their symbolic properties and relations.

My concern with this assumption comes back to claims made in the Emergence chapter. Reasoning conceptualizes phenomena. Thus, it is a process on which logic fundamentally applies. But when reasoning about AI as a modeler, we conceptualize the reasoning process itself. This converts the actual process into something that seems more amenable to description as a logical process than it really is. What we find is that, even though we perceive concepts with grounding, we can only discuss them using the explicit concepts they evoke. Of course, we cannot directly access the neural processes that underlie our reasoning, which have no explicit mapping to the emergent concepts and operations. The neural processes and the reasoning processes that emerge from them are not the same.¹⁸⁷ Therefore, it is unclear that humans process information with operations over explicit concepts, even if that is how we perceive it is done. We are conflating our own extrinsic emergent abstractions of the process, which is heuristic in that it compresses the process to enable easier external and internal (thinking) communication, with what could be a complex implicit process we cannot directly access.

The problem is not Logic itself but how logical assertions are constructed, and the representational properties required to construct them in ways that enable coping with the world in a similar manner to humans. If grounding matters for how information is processed, then one should not expect ERMs to generally be able to infer the same as we do using the concepts modeled. Yet, there is a widespread assumption that the explicit concepts are sufficient as model primitives, and the real task is to catalog their relationships and develop processes that can efficiently perform inference over them (i.e., symbol manipulation processes). For example, Douglas Lenat, founder of CyC (Lenat, Guha, et al. 1990), a large common sense ontology, is quoted as saying "*Intelligence is ten million rules.*" (Lenat May, 1988). However, as Cyc continues to add assertions to this day (7 million as of 2019, according to (Matthias 2019)), it is unclear that there is really a stopping point. One mistake here is the assumption that common sense is a finite set of discrete assertions. That is, the issue is not that there are too many rules to capture. What is missing are the underlying, subsymbolic properties of the human brain that not only activate, but construct any "symbolic network", as will be explained in Chapter 4. These types of representations used for ERMs are simply inappropriate for general Reasoning. Reasoning, even including **Common Sense, is generative, and the generation occurs subsymbolically. It is a process, not a set of explicit concepts and rules.**

C. Reasoning only requires explicit concepts but needs implicit understanding for sensory processing: A related assumption, one that appears popular with ERM modelers reconsidering

¹⁸⁷ As an example, the brain may perform Bayesian inference, as many suggest (Friston 2012), but whatever this Bayesian process is, it is not the same as directly computing probabilities over labels for explicit emergent concepts (e.g., probability that the label "cat" is true given label "pet" is true). The variables involved in neural Bayesian inference would be working with what are implicit variables for us. This matters if the concepts that do emerge from this process do not map effectively to what we as engineers would model within Bayesian models at the level of those explicit concepts (and thus would deviate from any AI models built to perform Bayesian inference directly at this level).

the challenges experienced with earlier ERMs, is that low-level perceptual tasks should be handled by feature learning methods (e.g., Deep Learning), but higher-level intelligence requires only symbol manipulation processes (i.e., explicit concepts to reason over). While this hybrid approach to AI has been around for decades, particularly in the research area of Cognitive Architectures (Kotseruba and Tsotsos 2020), it has been reinvigorated with the successes of Deep Learning largely being demonstrated on low-level intelligence tasks (i.e., image classification). I have no doubt that such hybrids can be beneficial for certain limited applications. Deep Learning methods have not demonstrated higher-level cognition. However, I see no reason to believe that higher-level intellectual processes would depend less on implicit dynamics than lower-level ones. Even if we discover that something other than (or beyond) the Deep Learning paradigm is needed, I do not subscribe to the notion that Reason/Cognition occurs at a level where logic and explicit variables are the modeling primitives. Just because we mentalize distinct concepts does not mean that our reasoning does not depend on subsymbolic context.¹⁸⁸

D. Logic over symbols is fundamental to intelligence, so it does not matter how important or skilled humans are at applying it: There is also the problem of how logic is applied by biological intelligence. The assumption that our reasoning tends to follow logical processes is suspect. Indeed, there is considerable evidence against such assumptions about Reasoning, including the claims about heuristics in earlier chapters and claims with supporting evidence in (Mercier and

¹⁸⁸ *Research: Graph Neural Networks (GNN) (Battaglia, et al. 2018) may offer a path to achieving higher level reasoning. While a benefit of GNNs is the ability to incorporate graph-based reasoning and explicit concepts as nodes and/or edges, the way learning occurs over the graphs is generally different than what occurs with ERMs, including, for example, vectorizing input graphs to train a vector space to learn graph embeddings. There are plenty of existing graphs to train on, which can be done in an unsupervised manner, similar to training with text data for word embeddings (i.e., predict connections or node labels or attributes within graphs to learn context for nodes and/or edges based on semantic distances from other nodes and/or edges). Existing ontologies, and structured graph or relational databases could provide graph data to augment training. Going further, one could fuse the knowledge base vectors with what has been learned about them in text corpora. This might provide an improved weighting process, as well as cover a large vector space span. It could also facilitate extracting relationships back onto the discrete graph (if useful). For example, a separate prediction model can be trained on vector operations using the known ontological relationships and then applying those operations on other vectors. Consider doing this with WordNet (Princeton University 2010), ConceptNet (Speer, Chin and Havasi 2017), and maybe even Cyc (Lenat, Guha, et al. 1990). Whether or not GNN's offer a unique path to AGI, I believe they can meet the criteria I lay out for grounding, although it depends on how particular models are architected. Given the flexibility in the general concept of GNNs, it is beyond the scope of this manuscript to assess specific limitations of architectures. Recent advances in this field are promising. Reinforcement Learning is another area where Reasoning is starting to become possible, such as through the use of Graph or Relational based methods (Džeroski, De Raedt and Driessens 2001, Jiang, et al. 2018, Zambaldi, et al. 2018). What these relational architectures demonstrate is the possibility that higher-order inference may indeed be possible without resorting to the content-less discrete symbols of ERMs. They exhibit some degree of grounding, and thus have the potential generalize to new scenarios without being brittle to environment changes that are irrelevant to the generalizable concepts learned.*

Sperber 2017).^{189,190} How skilled humans are at logic and the implications for AGI is a theoretical landmine, largely because humans have the capacity to recognize valid and invalid logic with enough clarity, and it would seem ridiculous to suggest purposely engineering AGI to be mediocre at logic (something computers can already compute with perfect validity on much more difficult problems than humans can solve). However, this is not the point being made here. The point is that logic is a means to an end in the evolutionary history that adapted intelligent agents able to reason, and we must get the end right for AGI if it is to be capable of human-level intelligence. As discussed in Chapter 2, there are heuristics we have adapted that determine how we apply logic in different contexts. These heuristics are not simply biases to excise when modeling human knowledge and intelligence. Intelligence, as we know it, *is* the application of these heuristics to generate concepts and how we use them for (implicit and explicit) inference and decision making. The intellectual medium for these heuristics is in the grounding of observations, the properties of which cannot be described explicitly.

E. Conceptual thinkers see the world as conceptual thinking: ERM's may be popular because of the typical educational backgrounds of AI researchers (i.e., Computer Science and/or Engineering). With computer programs, the variables mean something to the modeler, and the code determines how such variables, etc. will behave in different states (e.g., hovering mouse over location, X, Y). Many engineers therefore may see AI as a more flexible type of program (e.g., able to learn over time). The grand vision of such a perspective would be to build in all of ways for which we would expect humans to reason about the given explicit representational variables. The issue of grounding is not a secret, but it cannot be fully understood if one starts with the notion of computer programming in mind. I am confident in claiming that AGI is not yet

¹⁸⁹ Many studies show how relatively poor humans are at solving logic problems (Fantino and Stolarz-Fantino 2013, Kahneman 2011, Mercier and Sperber 2017, List of cognitive biases n.d.). Our Reasoning faculties appear to be biased in ways that were/are perhaps more effective in co-adapting with our particular social circumstances. A more fundamental point is made here, which is to question what it even means to reason over explicitly emergent concepts. Can we really decouple the concepts we adapted for effectiveness (rather than as objective constructs) with the processes we use to reason over them? Coping with the world clearly requires the ability to make seemingly consistent and coherent arguments given our concepts, but this does not entail that there is some objective answer about them that can be reached using logical over explicit emergent concepts. Logic is precise. The Scientific Method has also demonstrated the power of statistics to measure the certainty of inductive assertions. However, neither of these processes suggest that the constructs we develop, and how we reason with them, can be judged against explicit objective ways in which they "should" be reasoned with. For example, a model with concepts for Justice and Crime, etc. built using only on some set of explicit criteria (e.g., treat people equally regardless of their background, breaking the law) would likely produce some results that differ from what humans would infer given the same input conditions. This is because implicit inference differs from purely explicit inference, even if the former would be described, and thus conceptualized, similarly. It is not that Logic itself is wrong. It is the difficulty in creating explicit models that produce sufficiently similar inferences to humans that use grounded conceptual equivalents. Put more simply, we might be "correct" even if our described arguments are not (or *vice versa*), because we might not describe the full context, much of which might not even be precisely describable at the given emergent conceptual level.

¹⁹⁰ *Research: How must a neural structure be architected to learn to solve logic problems in a grounded way? How can memory be used without resorting to an artificial explicit memory store? Why are humans unable to process logic involving more than a few steps without stigmergic aid? Also useful would be to research how an ANN can understand what Logic itself means. An interesting project would be to study the relationship of understanding what logic means to how good ANNs are at solving logic problems.*

at the stage where we simply need to code it or integrate existing code. We still need new theories for how to even represent information in models.

F. ERMs are useful: The obvious reason why ERMs are developed, however, is that they have been useful in many domains. To state otherwise would be to negate all the Symbolic AI models ever built. It would be absurd to say that the decades of research produced no benefits. They are applied widely to provide insight to humans, both through the act of organizing our explicit knowledge and by performing logical and/or probabilistic computations that most humans cannot implement or would take too long for them to compute manually. Surely some of findings could serve as guideposts to advance aspects of AGI. The biggest contribution is likely the fact computer programming advanced in conjunction with Symbolic AI.

I believe ERMs share the same advantages and limitations of computer programming in general. Some ERM types add to these advantages the ability to compute uncertainty and update inferences given evidence to revise beliefs.¹⁹¹ Software has changed our world. However, there is a limit to what can be achieved. Programming, as it is typically done, cannot cope with Complex Systems in the way biological intelligence can. The promise of AGI is to enable software to move beyond this limit to perform tasks and solve problems at least as well as humans. My claim is that ERMs belong to the class of software programs we build today rather than the class of AGI that would achieve this promise. Importantly, this is a fundamental class distinction. That is, it is not an issue due to simply not building the correct ERMs, any more than one would expect standard software programming to achieve AGI.¹⁹² Effective Grounding is a major criterion for membership within a distinct class of models. The appearance of understanding in ERMs is the same illusion we find in standard programs. As discussed regarding the Chinese Room Argument above, the programmer uses their grounding of concepts to automate what programs output given inputs. The outputs can provide the illusion that the computer understands but only because the external programmer understood when developing the software.^{193,194}

Certainly, there are cases where correct answers can be deduced from explicit labels and rules. For example, a model could be used to deduce that, if queried for an animal that is a feline, orange with black stripes, and found in the jungle, then such a model could correctly return “tiger”, as long as all of those labels exist in the model with the appropriate relationships. Such operations are performed without any understanding of the variables operated upon, and the

¹⁹¹ The use of ‘belief’ here should be taken to reference human users of the ERM’s results. ERM’s have no intrinsic beliefs.

¹⁹² Given that essentially every form of AI is also software, the distinction I am making can easily be confused. By “standard” programming, I mean to differentiate it from code that engineers built intentionally as AI programs.

¹⁹³ To take the most obvious case, the output “Hello World” is not semantically grounded in the computer performing the program. The Hello World program constrains the computer to execute certain physical processes but are not grounded in any way similar to what it means for me to understand what the words mean or instructions for me to say or write them.

¹⁹⁴ If earlier claims about the eliminative/reductive nature of Concepts are correct, then one would expect that any program executions that humans do not know might appear to be intelligent but would be reduced to unintelligent causal sequences once fully understood. That this is true seems self-evident since we can generally assume causal sufficiency within the code for any software (minus any external inputs required, but which still must be ingested as bits), even if we have no access to that code.

operations must be fully determined and handcrafted by the human developer, but this does not matter if the inferences are correct. Moreover, Bayesian (and other) models can improve with evidence, which suggests some increasing consistency with the real world. What can we say about the domains that can be effectively modeled this way?¹⁹⁵

I tentatively believe ERMs work well in cases with following properties: First, they tend to be effective in domains that primarily involve deductive processes performed over concepts outside of common sense, such as Medicine and Tax Expertise. Since these domains are associated with higher-level cognition, their relative suitedness for being modeled with explicit concepts and processes, could be perceived as corroboration that higher-level reasoning can be captured with formal inference operations over explicit concepts. However, it could also be that novel technical concepts are naturally less grounded due to their lack of direct evolutionary pressures. In fact, since such specialized domains build off their own discoveries and concepts, newer concepts would be increasingly removed from any grounded heritage. If this is true, then humans may primarily apply formal reasoning because they have no other options. Thus, ERMs work in these cases because the domains themselves were constructed through formal reasoning. More accurately, ERMs can, in some cases, perform at least as well as humans in these domains because human models of the domains are themselves more like ERMs!¹⁹⁶

This leads to a more general understanding of where ERMs work well, which is where humans have already determined the criteria that variables must meet to be labeled as concepts, including criteria for changes in variables (i.e., the structural relations between variables). Consider a model that computes the probability that it rained last night given the grass is wet, the night sky was cloudy, and the sprinklers are off. Just as in the case above for specialized domains, before these processes can be computed, humans must first construct the variables, select which are relevant, determine which variables to relate to each other,¹⁹⁷ and determine what and how measured features will map to changes in variable values. In other words, even for this simple everyday domain, a human designs the experiment prior to the ERM inference performed.

¹⁹⁵ *Research: What makes ERMs effective models in some case compared to others given the grounding issues laid out in this chapter? An example is use of explicit representational Game Theory for Mutual Assured Destruction (Mutual_assured_destruction n.d.). What does it mean to claim that such problems are locally consistent? Is this the only explanation for why ERMs work when they do work? Are they more successful in cases where the emergent variables are relatively important (and then what would that mean?), or perhaps, cases where they are mostly independent of additional context? How is context defined here?*

¹⁹⁶ This is an overgeneralization for intelligence involved in specialized domains. As discussed later, there is a lot of implicit knowledge involved in, for example, coming up with the experiments involved in discoveries. However, note that ERMs are not good at these particular types of intellectual processes, further corroborating the claims in this chapter.

¹⁹⁷ I believe the need for AI itself to determine the variables and direct relationships may only be a practical requirement for grounding. The ability to learn the connections from data would not itself be a fundamental demonstration of grounding. However, the practical need to determine connections *a priori* does seem to demonstrate that biological intelligence involves some alternative to this explicit process to cope with the world. Since grounding emergent concepts seems to play a key role in this alternative, I include it in this list and others in the manuscript.

My claim is that much of human intelligence (i.e., what AGI must model) involves “designing the experiments”, including constructing the variables. At the very least, such a process goes well beyond what one would expect of low-level sensory processing. More importantly, however, there is no reason to believe that these processes always construct these “experiments” explicitly. While useful for checking explicit beliefs, examples like the probability of rain given wet grass is quite contrived. In most of our everyday mental lives, it is difficult to separate the implicit formation of models from the inference performed on them. Given that concepts are heuristic constructs, it may often be more effective to reason implicitly over non-conceptual percept properties themselves rather than only conceptual approximations of those nonconcepts. Indeed, the hypothesis that implicit processes may improve inferences or predictions seems to be exactly what feature learning methods have demonstrated (e.g., the success of Convolutional Neural Networks over ERMs for computer vision tasks). In other cases, we may find both implicit and (grounded) explicit inference are needed, where biases have led to increased effectiveness through our evolutionary history. ERM’s, however, are limited to features and variables explicitly understood by human modelers. In contrast, Feature Learning methods learn implicit variables that are maximally predictive or minimize uncertainty (e.g., novel features that increase predictive performance on diagnosing some disease). Again, even though humans seem to explicitly reason with explicit concepts, it is likely that the real inferential processes occur over features learned/evolved that are captured as grounding attributes for percepts.¹⁹⁸ This could be true both for nonconceptual processes, as well as actual inferential processes for which any evoked explicit concepts and explicit reasoning only approximate. The value of Grounding, then, is that models limited to only features and variables explicitly understood by human modelers can miss implicit variables that are more predictive or minimize the most amount of uncertainty. This is precisely what Feature Learning methods have increasingly demonstrated.

Two notes must be made about the claims directly above:

1. I am not claiming that less cognitive processes may be better for modeling intelligence. These nonconceptual processes are responsible for humans generating everything from new mathematical discoveries¹⁹⁹ to series of questions raised to work through complicated assertions.²⁰⁰ The phenomena of non-conceptual processes constructing the models we reason over is not any lower in cognitive level than explicit reasoning occurring over ERMs.²⁰¹

¹⁹⁸ *Research: This could be experimentally tested by measuring prediction accuracy, F_1 scores (F-score n.d.), etc., of humans over ERMs, while controlling for confounding factors such as human calculation errors.*

¹⁹⁹ That is, nonconceptual processes are responsible *in some way* for to enable such insights. If we knew how exactly, then we might already have built AGI. How such processes must be modeled requires further research. I speculate that claims made throughout this manuscript will help get there, but we certainly are not there yet.

²⁰⁰ Of course, questioning involves explicit concepts. The point here is that questions could be generated without explicit reasoning.

²⁰¹ If someone wishes to arbitrarily define explicit reasoning over explicit concepts as “high level intelligence”, then the point being made is that these “high level” concepts have incredibly limited scope within biological intelligence and improving them is not simply a matter of adding additional concepts, relating them explicitly, and perform valid operations over them.

2. I am also not claiming that AI must use nonconcepts to modify what would otherwise be theoretically optimal modeling of uncertainty, use and updating of probabilities, and/or logical consequences *given* explicit models. If a valid ERM predicts something different than a human operating over the same explicit concepts, and one suspects human estimation error, rather than missing context in the ERM (that is, the human and ERM are not solving different problems), then the probabilistic model is the one to trust, just like a calculator should be trusted over a human for difficult division problems. Humans have evolved many cognitive biases. However, these are contrived examples (cognitive biases were adapted for a reason!). They only demonstrate their precision and accuracy with the operations they are provided. I am arguing that most of intelligence involves implicit inferences that *cannot be constructed as explicit models in the first place*.

H. Grounding must be possible in ERMs because, no matter how you model AI, the model must consist of symbols all the way down. An underlying assumption of ERMs is the Physical Symbol System Hypothesis, which asserts that:

"A physical symbol system has the necessary and sufficient means for general intelligent action." (Newell and Simon 1976).

If everything is symbols, then this hypothesis is self-evident. However, I do not believe everything counts as a symbol. To call something 'symbolic' is to assume a reference frame.²⁰² In Chapter 4, I will define how this might determine what is and is not symbolic, which describes how grounding can be defined through the formation of symbols from subsymbolic processes. A claim will be that, for something to be symbolic, it must be grounded. It follows, then, that, **if all symbols must be grounded in something, and grounding involves subsymbolic processes, then the bottom for symbols cannot be other symbols. A symbol would have to ultimately emerge from subsymbolic processes.** I will also propose that **Intelligence should include this process of determining the symbols, not just the process of operating on them.**^{203,204} Within a given reference frame, it is not symbols all way down.

²⁰² Once again, I believe the conflation between Percepts and Concepts rears its ugly head here. As observers, anything we conceptualize becomes symbolic. This means that any features or properties of a model would appear to meet the definition of symbols. However, in many cases, they are not symbols from the theoretical reference frame of the models themselves.

²⁰³ Unfortunately, making a claim for what Intelligence is demonstrates the challenge in using only explicit concepts for assertions. Intelligence seems to elude any explicitly defined criteria, and thus any attempt to model it as such may result in some behaviors or abilities seeming less (or more) intelligent than they are as assessed using the formal criteria. Regardless, redefining it is not important for the claims in this manuscript. This manuscript is about what is required to build *General* Intelligence. If Intelligence is defined in such a way that symbol manipulation can in some cases meet the criteria, then the claim for General Intelligence is that there is something else required for it than simply Intelligence. The concern here is putting ERMs under the same umbrella as human intelligent processes, despite them being different not only in degree but also in kind.

²⁰⁴ When we look closely at what ERMs do, we find that we use them like we use calculators or statistical tools to quantify results of scientific experiments. Intelligence could be defined to include such processes, but I doubt it is what most have in mind for AI. At the very least, I think AI requires that these methods generate their own "experiments" (i.e., true model induction), and we will see that this cannot be fully done with explicit reasoning and representation.

For now, I will just express that I reject the Physical Symbol System Hypothesis because there is a necessary and sufficient way to ground symbols that does not involve a full regress of symbols. Representing emergent concepts as labels to operate in some system would miss the real cognitive processes. Where it seems to work, e.g., in scientific research, Bayesian graphs, etc. it is because the cognitive processes have been done by the human modeler upfront, but this removes the need for grounding in the model, as well as the possibility of that model achieving AGI.²⁰⁵ This is why I claim that **any ERMs are Models, not Modelers. AGI, like agents with biological intelligence, must be able to model, not just be a model.**

To summarize this section, ERMs may be used for various reasons. One reason is that they can be valuable in the same way that computer programs, when programmed effectively, can do what we want them to do, and often better than humans due to their ability to compute quickly, precisely, and reliably. This is not different in kind to a calculator computing a sum or a statistical software package returning a statistical significance value. Like computer programs, ERMs can solve some problems. Moreover, when data is limited, one can often get better results than feature learning alternatives (e.g., ANNs). However, while ERMs model concepts, they cannot model the cognitive processes needed to ground concepts and use them like humans as is necessary for AGI.

Qualities, Attributes, and Concepts

Given the above issues, a question is if there is an effective way to represent knowledge in models. In this section, I first take a step back and describe properties of knowledge, which will be important when discussing modeling implications.

As stated earlier, **subjective experiences encompass objective experiences.** Objective experiences have certain properties (e.g., analytically valid, consistent) that set them apart from what we typically call subjective, but obviously they are only ever experienced subjectively and privately.²⁰⁶ **Subjective experience consists entirely of qualities.**²⁰⁷ Qualities are instances of attributes. **Concepts, then, are just collections of attributes, or qualities when describing a concept instance.** When explicitly defined, qualities are described with adjectives. **Implicit Qualities** are what adjectives conceptualize. They are phenomena for which we may experience but have not assigned identifiers. As subjective experiences, percepts are composed of implicit qualities. Qualities can be made explicit as instances of concepts. Qualities are specific attribute values. For example, height is an attribute, whereas tall or short are qualities. Thus, an explicit quality is equivalent to an adjective or attribute value.

Knowledge and Reasoning also come in implicit and explicit forms. Explicit Knowledge consists of concepts, while Explicit Reasoning is the process of operating over explicit knowledge, often given some stimulus (e.g., observed environment or mental state). As

²⁰⁵ Markov and Bayesian models at least capture uncertainty that may arise due to the use of only explicitly represented variables (observed or hidden). However, I would not describe impacts to probability distributions for explicit variables “grounding”, and they do not meet the grounding criteria proposed in the Information chapter.

²⁰⁶ “experienced subjectively” is redundant but used here for clarity.

²⁰⁷ Hence, the term “qualia” is fitting.

discussed, ERMs assume that explicit knowledge and reasoning are all that matter for AGI. I have claimed that not only are they insufficient, but they fundamentally rely on implicit processes. With these terms in mind, **the question for AI is what is an appropriate way to model knowledge to enable reasoning?** The following sections discuss some options.

Graphical Models

Many ERMs can be modeled as graphs. This involves using nodes to represent concepts as variables and edges as relationships between these variables. Nodes often have properties associated with them, which can be considered their qualities. Edges may have a weight for magnitude. I will refer to these as **Graphical Models**.²⁰⁸ Nodes are necessarily discrete entities. At a symbolic level, then, they encapsulate concepts as a set of attributes. When attributes are set, they can be described as qualities that define an instance of a concept.

Is it natural to model concepts as ontological primitives? Since knowledge is emergent, it likely makes no sense to discuss a “natural” model representation. However, we can still check possible assumptions made for issues. If concepts are only experienced (i.e., grounded) by their qualities, then it would seem that qualities would be ontologically primitive. Humans, of course, conceive objects. Even if perception began evolutionarily through the experience of two qualities (e.g., dangerous, not dangerous), as the repertoire of qualities experienced grows, it becomes effective to group some that are commonly experienced together. As discussed in Chapter 4, abstractions should compress message lengths to enable easier communication and possibly cognitive operations by applying more manageable symbols.²⁰⁹

Concepts are heuristic abstractions (e.g., nouns) of a set of qualities (which are themselves heuristic abstractions that can be conceptualized as adjectives²¹⁰). It is thus the qualities experienced that determine the “essence” of concepts and concept instances, rather than the concepts themselves having some essence beyond the qualities that ground them. Evidence for the qualities determining concepts is easy to find. In “Surfaces and Essences”, Hofstadter and Sander provide many examples of objects having different “identities” depending on context (Hofstadter and Sander 2013). For example, a historic airplane at an airport could be an airplane, statue, or symbol. There are no clear cut “things in themselves” given that we categorize things solely on the qualities we perceive with them. Each of the “correct” identities, e.g., for a plane, depends on what subset of possible qualities are considered at a given point in time. Whether we believe it is the same object in each case only depends on what qualities we select as the criteria for deciding, and whether those persist across the different quality sets we are comparing. For example, whether it is an airplane or statue depends on the qualities that we assign to it, and whether it is considered the same thing in both cases depends on whether we

²⁰⁸ Not all ERMs are best represented as Graphical Models. Also, graphical model nodes and edges do not have to be explicit (e.g., ANNs), but it is the explicit models that are of interest in this section.

²⁰⁹ *Research: If nouns are grouped adjectives, consider if/how this might modify Chomsky's grammar production rules (Chomsky 1957). Specifically, this replaces the noun in each production rule with a set of adjectives. Consider if and what all the other variables (e.g., verbs) can be reduced to. Is the resultant universal grammar more generalizable to include examples where actual language use breaks formal production rules?*

²¹⁰ We will see how adjectives can be grounded as concepts in Figure 2 below, which converts them into nouns.

believe other qualities, such as being in the same location, made of the same materials, set the criteria for it being the same thing. Thus, at least epistemologically, there is nothing beyond the qualities evoked in the observer that objectively determine our concepts.

This matters for AI that attempts to use explicit concepts as primitives. There is no reason to assume concepts can form a logically consistent model of the world. How qualities evoke concepts is not governed by the logic we apply at the level of concepts as things in themselves. In graphs, context is provided via edges to other labelled vertices, and the vertices also have a limited set of labeled qualities. Moreover, many qualities are continuous phenomena, where different blends of them create what may be observed as distinct concepts. If concepts emerge from them, then AI requires a continuous space from which concepts emerge.²¹¹ Indeed, I claim that this is one of the criteria for grounding. Labeling a *node* = *dog* with the quality *weight* = 100 *lbs.* has no grounding for animal or weight; consider how you experience the quality of dog and 100lbs vs. dog and 10lbs. The, perhaps implicit, experience differs between them in a way that is much richer than what differs for the computer (a change in one ASCII character). The grounding that biological intelligent processes apply was adapted to facilitate certain expectations and effects from our external world. Without that grounding, one cannot expect models to lead to the same expectations and effects, except under rare circumstances, as described elsewhere in this chapter.

Another issue with Graphical Models is how large they must be to capture relevant contextual information in explicit form. The world of concepts is not only vast and highly interconnected but has the potential for infinite concepts and connections because they are formed generatively from the continuous spaces of qualities. This suggests that there is something unnatural about modeling concepts and their properties in discrete sets of nodes and edges, since it would never scale for AGI.

Conceptual Spaces

There is an alternative way to represent and reason over knowledge that is consistent with claims above. Again, a limitation of graphical models is their discrete representation of emergent concepts. This discretization not only limits the number of modeled concepts but loses meaningful context across nodes based on their sets of qualities. Addressing this not only requires a continuous space to model qualities but for the continuous space to express meaningful distances between any sets of qualities. That is, it must be, or at least approximate, a metric space. Peter Gärdenfors (Gärdenfors 2014) provides an elegant way to address these requirements. This is done by representing attributes as dimensions to create a hyperdimensional space within which concepts are positioned. This is called a **Conceptual Space**.²¹² They will be leveraged for many of the claims made throughout this manuscript. Unlike

²¹¹ Although it is difficult to explain concisely, we will also see in a later chapter that this does not invalidate the possibility of an even more foundational graph or network to form the continuous space for concepts, but only if the lower-level graph is not modeled as consisting of symbols from the reference frame of the continuous space. As I stated earlier, reference frames are critical to understanding how to ground symbols.

²¹² Given my own distinction between Percepts and Concepts, I believe the term “Conceptual” may be too narrow for what these models capture. Anything we consider externally for a given model, however, would be

graphical models, the distance between concepts can be meaningful, creating an infinite and potentially bounded space that efficiently captures all possible concepts based on their attributes (e.g., no need for a modeler to specify attributes for each separate concept) (Figure 2).

Figure 2. A small example of a conceptual space to represent the concept 'apple'. Figure taken from (Banaee and Loutfi 2014).

Conceptual Spaces provide numerous advantages over explicit Graphical Models, including how new concepts can be generated. The simplest possible case is a number line. Any point selected will be a new number since the line is continuous and meaningful (i.e., the distances between all possible points captures their different magnitudes). A simple multi-dimensional example is a Color Space (Figure 2, right), where the attributes are brightness, saturation, and hue. Here, any point in the space represents a unique color.²¹³ A discrete model for knowledge cannot generate new concepts this way. This also addresses an issue of scaling graphical models, since the hyperdimensional space can cover a continuous space of possible concepts and relationships. Because the distances are meaningful, vector operations can be performed without having to model each relationship between all concepts explicitly. While a graphical model can in theory capture distances as parameter values, there is no clear way to extract new nodes in between known ones meaningfully.²¹⁴

A Conceptual Space can capture many, if not all, of the representational capabilities of Graphical Models. For example, if concepts can have a range of values, either as uncertainty or as attributes that define the concept, then they would take up regions of qualities, creating a subspace. More speculatively, uncertainty could also potentially be captured, such as by a conditional “fuzziness” based on a given probability distribution for dimension positions of concepts, creating a conceptual subspace, with hypervolume proportional to certainty. For

conceptualized. For sake of simplicity, my use of the term ‘Concept’ in this section will be in the way it is used in the Conceptual Spaces literature, even though ‘Percept’ may be better depending on the reference frame (i.e., what is a concept to us as observers of the space is a percept to intelligence with the space internalized). I discuss this point later in this section.

²¹³ Note that Conceptual Spaces capture semantics, and yet this Color Space requires no natural language explicit in the model. Semantic properties are thus not limited to natural language, even though that is the common domain in which semantics are discussed.

²¹⁴ Composition rules could be implemented to form new nodes in some graphical models, but the production of such nodes or edges is arbitrary except for the infinitesimal fraction constrained by specific handcrafted rules based on external knowledge. That is, with enough handcrafted rules, some models can begin to resemble meaningful modeling but only for very narrow parts of certain domains. My claims in Chapter 4 are that any such resemblance is not truly modeling on the part of the AI, as the lack of ability to scale this way demonstrates.

example, if we do not know if an animal on the street is a cat or dog, then this could create a probability distribution over the subspace for those two animal types. This creates conceptual overlap between the idea of uncertainty and higher-level concept classes, as both relate to the hypervolume in the Conceptual Space.²¹⁵ Conceptual Spaces can also capture hierarchically abstracted concepts by changing the size of the concept subspace's hypervolume. Larger hypervolumes add attributes and/or increase attribute ranges to capture higher-level concepts. For example, the concept mammal could have a larger hypervolume that contains the hypervolume for the concept cat.²¹⁶

Relations in Conceptual Spaces can be implicitly reasoned over through vector operations. This enables similar forms of inference provided by Graphical Models but with the unique ability to apply them to countless concepts. For example, an analogy-making operation *A is to B as C is to D* = $\vec{B} - \vec{A} + \vec{C} = \vec{D}$, where the arrows represent that the variables are vectors, can be done at any position in the Conceptual Space to derive new analogies.²¹⁷ Again, this is made possible by the distances within the continuous vector space providing the semantics, and therefore the grounding evoked and associated with symbols we receive and generate. A Graphical Model with discrete explicit concepts would need to apply the same analogy to each concept separately.

Again, the potential concepts that can be evoked from a Conceptual Space are countless whereas, with Graphical Models, every concept and their relations must be separately added to

²¹⁵ Research: If concepts are subspaces with size depending on the dimension values the concept can take, then in what cases, if any, does it make sense for lengths across certain dimensions to represent uncertainty? The uncertainty would have to be over only those attributes that the given concept instance might have. How would the probability distribution be captured (i.e., that some values are more likely than others)? Could probability/uncertainty be constituted by missing relationships between concepts and/or missing dimensions. This would require that a given conceptual model can have patchiness or vagueness within the continuous space. What underlying processes would create such differences within the vector space (perhaps an underlying ANN that has trouble "activating" certain vector positions?) If there were somehow missing dimensions, then this could increase uncertainty or abstractness for things that differ in the missing dimensions' qualities. Also, is an assumption of orthogonal dimensions (i.e., qualities) correct or do some qualities naturally correlate? Consider if there is any intuitive relation between higher-level percepts and uncertainty. Would this be more similar or less for how we model the world? Consider if uncertainty should instead be its own dimension. This allows single vector activations but seems like it would capture the perceived uncertainty rather than a true uncertainty that would reduce specificity over possible concepts.

²¹⁶ This notion of larger subspaces for higher-level concepts makes sense from the perspective of trying to capture knowledge logically. However, cognitively-speaking, I suspect that higher-level concepts would often have *different* quality types than some of its members. For example, one might in part ground the concept of Earth in a city-like quality but not so for the concept of Planets. It is technically a quality for Planets since at least one planet has cities, but there is no reason to assume that a given agent's conceptual space is this literal. As we will see, Effectiveness constrains the actual Conceptual Spaces found in biological intelligence. Note that there must be a way to still comprehend that a city-like quality is logical for Planets when the fact is pointed out (as is done in this footnote). Presumably, this involves some vector operations and/or subspace combinations, rather than requiring some separate logical model. It is interesting to reflect on how the grounding underlying different perspectives for the "same" concept changes in these different circumstances.

²¹⁷ Indeed, comparing results from analogies seems like a useful way to determine how close a Conceptual Space is to our own intelligence. If the Conceptual Space performs well over the majority of analogies we use, then it is likely pretty close to reaching general intelligence. Given that there are infinite possible analogies that could be triggered from a Conceptual Space, it is unclear how much we should expect to understand or agree with randomly chosen analogies, but one would hope that, once an effective Conceptual Space has been built, many of these would be understandable and insightful.

the model. This is not just a practical issue. No discrete model would ever be able to capture all the knowledge of humans because the way information is modeled requires the ability to generate knowledge, including the ability to generate concepts with variations on grounding never before experienced.²¹⁸ As a generative process, it requires information to be represented in a more open sense. Conceptual Spaces provide such a representation, where new information can impact the distances between existing information, thus potentially modifying local semantic properties through global context. The implicit vector space properties constrain the properties of each possible concept, but it is open due to its continuous nature.

Moreover, with infinite possible new concepts, there is practically boundless potential for novelty. Our insights build off of existing grounding. This can involve reasoning with explicit concepts, but it seems more plausible that it is the grounding influencing and influenced through such reasoning processes that evokes the understanding experienced with, and core to, insights. This is thus important for comprehension and creativity. A metric space, such that distances modify semantic properties, is necessary for grounding their generation.

More informally, Conceptual Spaces just seem more natural for modeling cognition. Since concepts are emergent, I do not believe they can be grounded simply by operating over only explicit qualities.^{219,220} There is something about each quality that a label invokes in us that the label does not itself contain, leading to an infinite regress of absent qualities.

There is at least one potential challenge with Conceptual Spaces: They seem impossible to manually construct to capture most human concepts. There are approximately 10,000 adjectives in the English language. Humans have difficulty comprehending anything beyond three dimensions, yet a Conceptual Space that covers many of the concepts we use would have many more. Moreover, there is no reason to believe that our own quality dimensions map one-to-one with the explicit qualities we have emerged as concepts. In fact, concepts for qualities should be derived from the Conceptual Space itself. While it is possible in some cases, that they map precisely to explicitly understandable dimensions, such as saturation and hue, it seems likely that many dimensions would be implicit. If concepts require implicit qualities, then this turns a difficult engineering task to something impossible to manually engineer.

Fortunately, recent Machine Learning advances seem to be demonstrating a scalable way to construct such dimensions using data and without needing to assume an explicit quality that each must represent. While it is unclear that this a natural way to construct Conceptual Spaces, the

²¹⁸ Indeed, I claim later that this may be why our thoughts are often experienced as being unique, even if any associated concepts experienced are known. ERMs cannot create new emergent experiences, yet our own knowledge is clearly generative in this way.

²¹⁹ *Research: Combining two concepts could in theory change the grounding of the previously separate concepts, making their new subspace more than the sum of the separate subspaces. Is this consistent with what Conceptual Spaces would represent? Does this mean that there is information that cannot be explicitly mapped to the model's assumed representation? Note that such combinations irreversibly (in a Formal Causal sense (Four causes n.d.)) create new grounded properties. Determine how this lossiness might be measured in Conceptual Spaces. This is consistent with the adaptive value of concepts as a means to compress information into smaller message lengths. Can these processes be measured in humans as they learn to mentally chunk concepts (Chunking_(psychology) n.d.)? Is there a way to empirically determine this for AI (e.g., by having AI evolve the need and ability to learn and group concepts and somehow measuring grounding differences?) Consider what changing scales (i.e., resolution of dimensions) might show, if anything, including any potential relation to emergence levels.*

²²⁰ *Research: Consider how attribute values (qualities) in Conceptual Spaces relate to specific types of adjectives, etc.*

results have been impressive. I turn to these Neural Embedding / Semantic Vector Space methods next.²²¹

Neural Embedding / Semantic Vector Spaces²²²

Considerable research has been recently produced on what are called **Semantic Vector Spaces** that are formed through Neural Embedding methods. Although not synonymous, they seem to capture many of the core attributes of Conceptual Spaces, such as representing concepts as positions within a vector space that approximates a metric space. Indeed, it has been found that some vector operations result in the types of results one would expect from Conceptual Spaces.²²³

Rather than each dimension of the space being labeled as an explicit attribute, they represent one neuron (i.e., node) each from a given trained ANN hidden layer, with synaptic weights learned from data (e.g., corpora of text). The neurons across the given hidden layer thus are what constructs the Semantic Vector Space (Mikolov, et al. 2013). This in theory addresses the challenge of engineering implicit attributes. Although ANNs hardly resemble biological NNs, the fact that these Semantic Vector Spaces can emerge from them, and learn from observations, also suggests a natural way for them to be modeled in the brain.²²⁴ I am not aware of a similarly elegant solution for how concepts might arise from biological NNs.

What is the result of these implementations? So far, it has been state-of-the-art Natural Language Processing, among other breakthroughs (Tum 2020).²²⁵ The requisite complexity and lack of explicit rules enable what no one has been able to handcraft in the past. I believe what is

²²¹ *Research: Consider Conceptual Spaces in relation to processes as discussed in Process Philosophy* (Whitehead, Griffin and Sherburne 1979). *What theories or insights can be built from studying vector dynamics?*

²²² The term 'Semantic Vector Space' does not readily differentiate itself from the term 'Conceptual Space'. I will therefore interchangeably use 'Semantic Vector Space' and 'Neural Embedding Space', depending on the need to distinguish the methods from Conceptual Spaces.

²²³ A common example used is $\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} \approx \overrightarrow{Queen}$.

²²⁴ ANNs are themselves edges and vertices. This further suggests that the problem is not inherently due to graph representations, but due to the assumption that knowledge (and operations on that knowledge) can be explicitly and discretely modeled; something the neural embedding methods do not assume. Again, this points to how AGI research still needs to address how to even represent information.

²²⁵ **Despite these breakthroughs, training from only text data, no matter the amount, will not ground the concepts contained like how we ground them (not even close). Despite this, I do claim that these models can ground the text in some form, and that this is what makes them work as well as they do. That this form is but a shadow of human-like grounding means they cannot achieve AGI on the current path. However, that the models meet the criteria I will claim for grounding at all is why it matters for the future of AGI. This must be understood as the argument being made in this manuscript.**

being demonstrated at least hints at how grounding can be modeled and how that grounding enables inference.^{226,227}

Skeptics of these approaches usually point to the vast data and processing requirements, which are necessary for sufficient semantic coverage, as if these methods are using some performance enhancing drug. This is certainly a major limitation, and one that must be addressed to be on a path towards AGI. However, whether or not the data requirements become a permanent barrier, I believe it misses the key insight these models are already providing. The models are capturing something fundamentally missing in ERMs. This includes the value in having a model with the representational capacity to learn a continuous metric space. This seems like the only option to infer things not explicitly populated into the model itself. The underlying neural network can emerge the symbols for concepts based on their position in the continuous metric space. This, for example, is required to extend categories for things to include those never previously observed by the model. If this is not grounding, then I do not know what is. I will elaborate on this throughout the manuscript.

While these are providing breakthroughs in Natural Language Processing (and other) challenges, there are other issues to address. A major one is the limited interaction these models with their environments in current research. Grounding for biological intelligence has been strongly shaped by evolutionary pressures. I can therefore only imagine how much better these models could be with the rich interactions described in Chapter 2 above. For example, our own evolutionary history must explain why we group certain sensations into e.g., an emergent concept of a couch that is separate from the rest of the environment. This is not completely arbitrary. Lower-level mechanisms and laws must exhibit some degree of regularity to evoke an effective emergent concept, but other groupings could emerge with alternative evolutionary

²²⁶ I am optimistic that inferencing capabilities in these Semantic Vector Spaces will increase in sophistication in the coming years. In fact, a bolder but more tentative prediction is that research attempting to populate traditional ontologies using semantic vector spaces will end up having minimal long-term value. End-to-end embeddings retain grounding that is lost when only used to feed a coupled but distinct ERM of any sort. There are benefits for human users of the models, such as explainability of inferences, but this does not necessarily extend to the AI being modeled itself. That said, humans experience symbols as internal communication (either using them for thinking or experiencing them as emerging from the subsymbolic thinking), and it may turn out to be necessary for AGI as well.

²²⁷ *Research: Beyond determining a set of inferences to explore, consider how known cognitive scientific knowledge and heuristics can be defined over a vector space, including Pearl's Causal Do-Calculus (Pearl and Mackenzie 2018). This is more natural than current approaches to scalable AGI, which simply outputs classes from perception-level deep learning and feeds those as class symbols to explicit reasoning components with none of the grounding preserved. Instead, Symbolic AI methods might be used as heuristic operations over a vector space that persist across levels. How well do the vector operations scale? It is hardly valuable if we must build millions of them. One option to consider is having a separate neural network learn the best operations (the ANN reasoner) if there is labeled data, or at least some sort of cost function to use. This could also be an evolutionary or reinforcement learning algorithm.*

pressures (i.e., the patterns “out there” are chosen based on some adaptive heuristic).^{228,229} As I will discuss in later chapters, Semantic Vector Space models could also be shaped by effectiveness through the application of some driver, such as a reward or fitness function. Such an approach would in theory address the needs described in Chapter 2.^{230,231,232}

Perceptions and Conceptions in Conceptual Space Models

Now that Conceptual Spaces have been described (including as approximated through neural embedding methods), I consider them with respect to the nature of percepts. Let us first note that the representational model is called a “Conceptual” Space. Where do percepts fit into this model? Based on my earlier claims, whenever we discuss a property within the space (e.g., a vector or hypervolume), the discussion about it forms a concept. Also note that I avoid the assumption that percepts are something that occurs prior to some activation in the space (e.g., a neural layer upstream of the hidden layer representing the Semantic Vector Space). This would be easy to assume due to the popularization of Semantic Vector Spaces being associated with training on text data. One might suppose that some lower-level sensory processing would occur prior to Semantic Vector Space activation. However, these spaces can be trained just as naturally on imagery and other types of data (Beaumont 2020, Lynch, et al. 2020). Even though there are no explicit terms developed in these nonlinguistic spaces, it is appropriate to describe what they model as Semantics. For these reasons, I tentatively claim that positions in the Semantic Vector Space really represent percepts for the model (i.e., when considered from the theoretical perspective of the model itself). However, in an advanced model at least, there will be points that are percepts of the conceptual type. If so, then we should be able to analyze these spaces and find positions that capture aboutness, perhaps even of other positions in the vector space. Thus, a point in the color space would be grounded as a color percept but would not be about the color. To be about the color requires something else.²³³ Again, this may seem difficult to believe because, were we to analyze any of the vector in a Semantic Vector Space, they are converted to being about the properties associated with that vector. My claim is that this would not be so if

²²⁸ *Research: Determine how an evolutionary model could be used to evolve effective group-building rules (e.g., the formation of the percept/concept of couches) and by what criteria those rules could be associated with an implicit mapping formed across emergent levels for agents. Evolutionary algorithms should be able adapt specific pattern formation/detection emphases, and the agents must evolve to be sufficiently complex for effective, robust, and detectable mapping rules. Different evolution runs should be used to assess the effects of historical contingencies in the evolved maps. The grouping that develops should correspond to the types of processes that are most effective, as is demonstrated with the Empirical Theory of Perception (Purves and Lotto 2003).*

²²⁹ *Research: How do “laws” between higher-level concepts relate to lower-level “laws”? Can the relationship be described?*

²³⁰ **Although out of scope for this manuscript, I believe that AGI will actually need something more fundamental than predefined reward, fitness, or any other optimization functions. A subsequent manuscript is in process that proposes promising research avenues to emerge optimization processes from more fundamental dynamics.**

²³¹ Given a sufficiently effective objective, this would also enable AI to be considered creative by generating completely new concepts with functional value.

²³² *Research: Develop a study that could determine how close Semantic Vector Spaces dimensions are to various explicit attributes.*

²³³ ***Research: What would this “something else” be? Would it be something like holographic properties where certain dimensions contain all the information of the higher dimensional vector space? Is there a simpler way to envision vectors with meta-vector attributes? Determine what type of architecture would be required and how to train and test for such properties.***

were we able to assume that the vector space model experiences those vectors and that we could take its perspective.²³⁴

While this is important to consider, I will not belabor the point by replacing the term ‘Concept’ when discussing Conceptual Spaces. There is much more to discuss in this manuscript, and the distinction could distract from other points. All one must keep in mind is that some of the Conceptual Space vectors discussed are better described as Percept vectors when considered from the reference frame of the system itself (i.e., the AI or biological agent with the Vector Space). Instances of percepts (including instances of the concept variety) are vectors.

Mapping and Model Types

With the properties of these different ways to represent knowledge described, we now consider them in relation to the challenge of mapping across emergence levels. The challenges with ERMs were lack of grounding and lack of meaningful distances between concepts.

The relationship to mapping and implicit properties should be clear at this stage. **Emergent phenomena map to lower-level mechanisms in the sense that they are caused by the mechanisms (and can be eliminated by them in conceptual form), but this mapping is implicit. That is, there is no description for *how* the causes emerge these phenomena.** We know what atomic elements emerge water, but there is no description for the emergent properties of water. As discussed, the Description Gap means that there is no way to explicitly map from mechanisms to emergent concepts because emergent concepts and maps are described at an emergent level without a means to reduce it to mechanisms. How does this relate to grounding? The question is whether capturing implicit properties enables more effective mapping. While we cannot know what implicit properties comprehensively map to mechanism levels, we do know that they matter for inference and prediction. Feature Learning methods have established their ability to learn features that provide much higher performance ceilings for inference and prediction vs. ERMs. ERMs require the symbol grounding of the human modelers to structure them but themselves have no understanding of the features and variables with which they are labeled. In contrast, some feature learning methods seem to ground features and variables that humans cannot cleanly map to explicit features and variables. Such models directly observe features and abstract them into grounded emergent concepts.

This is admittedly vague at this stage. Grounding must be better defined, and such a definition should demonstrate why these ERMs are less grounded than, say, an ANN representing the same concept but with subsymbolic properties captured as a distributed representation.²³⁵ Later I will describe how Semantic and Effective Processes that make Information possible from the perspective of a model fulfill this role.

The upshot for ERMs is that starting with explicit concepts without modeling any of the implicit grounding that gives rise to them cannot learn an effective map to lower-level mechanisms, and without grounding similar to how we ground these concepts, we cannot expect the ERMs to capture a comprehensive model of human Knowledge or Reasoning. Therefore, we should not expect them to produce results (e.g., predict, infer) sufficiently similar to human

²³⁴ This is purely theoretical. As I said at the beginning of this chapter, I cannot make any claims about whether these or any other models really do experience anything (i.e., have intrinsic emergence).

²³⁵ Ideally, we would have already been able to compare human mental models to ERMs, but we lack the tools and non-invasive access needed to measure this in humans with sufficient precision.

inference and prediction results, as we would expect for AGI when given similar data. Agents with biological intelligence do not use just the explicit labels for inference or reason. They apply implicit grounding properties that evoke (and thus ground) the concepts that emerge from inferencing and reasoning processes.

A challenge that I have already raised is how grounding through feature learning methods could emerge results similar to humans. For one, current methods are applied to data that often do not resemble what biological agents receive, and we should therefore expect deviations in their inferences and predictions compared to humans, who interact with mechanisms at the level of physical processes. Of course, we do not have the knowhow or computational resources to build in fundamental physics and initial conditions to produce implicit maps that might more closely align with those learned by humans. AGI requires implementing grounding as effectively as humans do. Ideally, there are enough common constraints in the world such that data received by AGI models can be used to emerge maps that fall near or even within the range of those adapted by intelligent biological agents.²³⁶

In Chapter 1, I claimed that explanations for emergent concepts are eliminated with the concepts at the level of mechanisms, leaving an unbridgeable Description Gap between mechanisms and the concepts humans construct to explain the world. Grounding it is not just a property of the message by itself. It depends on the recipient's model. **To ground a message is to associate it with some observer's existing knowledge. As we will see, different types of processes relate to grounding and therefore mapping concepts, including Semantic Processes, which associate messages with an observer's implicit and explicit knowledge of the world. Therefore, even though we cannot model explicit maps (e.g., cannot describe how lower-level mechanisms are associated with emergent concepts), we can have AI learn to ground symbols it observes with the same general types of processes we use.** The discussion in this and other chapters help to understand the requirements, as well as potential measures.

If rich information and interaction are required, then what specific properties must an AI model meet to be considered grounded? I will argue later that Semantic Processes provide appropriate measures for grounding, and Effective Processes are critical for forming useful Semantic Processes. This suggests that the path towards AGI must in part be guided by how we measure these processes.

²³⁶ Humans also evolved their effective maps without such knowledge. Even if limited, the commonalities across humans' models are sufficient for the levels of social coordination we achieve. This would not be possible if concepts communicated were not grounded somewhat similarly across humans.

Chapter 4: Information, Semantics, and Effectiveness

Introduction

So far, I have been describing some of the challenges with emergence and grounding with respect to AI. The ultimate objective of this chapter is to determine how we can measure important concepts associated with grounding. This chapter provides what I believe are important claims and considerations with foundational implications for AI and the notion of Information itself. The claims and questions raised relate symbol grounding to information, semantics, and effectiveness. Future research will hopefully take these into account and converge on the most appropriate and valuable formalizations not yet captured. If the claims made help address grounding and its implications, then they will in turn help to advance AI and AGI.

Deep Learning has addressed some but not all AGI challenges. This rightfully leads to concerns raised about their, at least near term, limitations. However, as I have claimed throughout this manuscript, it also raises concerns about going back to methods that also did not address fundamental challenges, or what I have grouped together under them term, ‘Explicit Representational Models’ (ERM). These challenges include the relation between Information and Emergence, as well as the nature of Semantics.

Shannon Information Introduction

There are different definitions of Information, however, I will largely focus on the most common one, which was defined by Claude Shannon (Shannon 1948).²³⁷ **Shannon Information** can be described informally as a measure of change in uncertainty about some variable X given an observation. Typically, a probability distribution for X is defined and the observation is a value for X from that distribution. A common example is a coin toss. For a fair coin, landing on heads has a 50% probability as does landing on tails (i.e., a uniform distribution). This probability distribution tells us the uncertainty (i.e., Shannon entropy) of what value a coin toss will take, as well as the amount of Shannon Information expected to be gained by observing the result of a coin toss. Shannon Entropy is smaller for a biased coin because one is proportionally more certain about the result of a coin toss. Shannon Entropy is greater if tossing two fair coins rather than just one fair coin because there are four equiprobable results instead of two, and thus less certainty.

While these examples provide an intuition about Shannon Information, the emphasis in this chapter will be on the role of **Microstates** and **Macrostates** in measuring Information, which is more directly associated with the role of Information in Cognition. Imagine that, rather than receiving the actual results of the toss of a pair of six-sided dice, you are told the outcome sum. For example, observing that the sum of two fair six-sided dice equals five (i.e., macrostate) decreases the uncertainty about what the possible values are for each of the die (i.e., microstate).

²³⁷ I do not provide an in-depth overview of Shannon Information Theory or other existing Information Theories since other resources exist at any level of specificity and comprehensiveness desired. What I do describe should be sufficient for understanding the claims in this chapter.

After the roll but before the sum was known, there are $6^2 = 36$ possible microstates.²³⁸ Shannon Entropy²³⁹ is measured as:

$$H(X) = -\sum_{i=1}^n P(x_i) \log_b P(x_i), \quad (1)$$

where x_i is the probability of each possible microstate (e.g., each possible value pair). When the logarithm base, b is two, the measure is in bits. Since each microstate is equiprobable, the Shannon entropy after the dice are tossed but before being told the sum equals:

$$H(X) = -\sum_{i=1}^{36} P(x_i) \log_2 P(x_i) = 5.17 \text{ bits}. \quad (2)$$

The intuition behind this value is that it would take on average 5.17 binary (i.e., true/false) questions to know a microstate rolled. This is an average because the dice values are randomly sampled, and we must therefore account for all of the possible values that would be sampled over infinite trials. Entropy is therefore also associated with the expected information gain.

Now we consider what happens after being provided the sum value of the toss. Since we are still interested in determining the microstate, the sum value can be thought of as a macrostate. Consider the Shannon Entropy $H(X)$ of the observation of a sum of 5. Four of the 36 microstates sum to 5: [1,4], [4,1], [2,3], [3,2], each with probability = 0.25. The Shannon Entropy given this macrostate is:

$$H(X) = -\sum_{i=1}^4 P(x_i) \log_2 P(x_i) = 2.00 \text{ bits}. \quad (3)$$

Thus, it would require on average two binary questions to determine the true microstate. If instead, the sum observed is 3, then $H(X)$ would equal 1.00 bits because only one binary question is required to determine which of the two possible microstates, [1,2], [2,1] underlies the macrostate ($sum = 3$). Accordingly, a sum of 3 has higher Shannon Information Gain than a sum of 5 for this example because it results in a greater reduction in uncertainty about the underlying microstate.

Generally, we can think of Shannon Entropy as a measure of the uncertainty of a microstate, and Shannon Information as how much a message reduces the number of possible microstates that could have produced the symbol(s) in the message received (e.g., the sum of the dice), where the distribution of possible microstates is the macrostate.²⁴⁰ The macrostate can thought of as the observer's (i.e., receiver's) model of the environment (Figure 3). In the dice example, the model is simply the distribution of individual dice values that equal the known dice toss sum. However, we will see that for biological intelligence, knowledge can be much richer for certain agents and environments, creating more complex macrostates.

²³⁸ In this case, the base = 6 is the number of possible values for a single die and the exponent = 2 is the number of dice rolled and summed.

²³⁹ I will just call this "entropy" where there is no confusion with other possible uses of the term.

²⁴⁰ I will tend to refer to "Shannon Information" over "Shannon Entropy" for this chapter, since Information is easier to understand (entropy is the expected amount of information). They are derived from the same equations, so any differences between the two do not affect the claims in this chapter.

Shannon Information Requires Predetermined Symbols

A key point in this chapter is to note that Shannon Information can only be measured against a macrostate and microstate consisting of a set of symbols, which can be thought of as the indices for those states.²⁴¹ The macrostate is the receiver's model of something. Shannon Information can be measured for any symbol set, but the symbols must first be determined. For dice, the symbols are counts of black dots of a certain size on each face of the dice. If a receiver had decided that the symbols were instead the number of rotational symmetries for each face of the dice, then the Shannon Entropy could differ.²⁴² One can change the symbols for the microstate and macrostate or both. That is, the microstate could be the number of rotational symmetries for each face counted over two dice, while the macrostate could be the sum of those given sampled rolls. The microstate could instead be the pair of individual dot counts as before, but the macrostate could be the sum of rotational symmetries.

For these sets, the symbols are numbers, but they *mean* different things. That meaning must be determined by the receiver of the symbols.²⁴³ Symbols are received in **messages** (Figure 3).²⁴⁴

²⁴¹ "Symbol indices" (or just "indices") can be thought of as what are found in model equations, often under summation or integration operators. An index contains the variables from which a given model is composed.

²⁴² Standard dice have faces where one, four, and five each have 4 rotational symmetries, one for each 90° rotation, whereas two, three, and six each have 2 rotational symmetries. For six-sided dice, the Shannon Entropy for the sum of one or two rolled dice is approximately 2.6 and 3.3, respectively for using the count of dots as the symbol index, while it is 1.0 and 1.5 for using the count of rotational symmetries. Note that what is fair for dice for using the dots is biased for using rotational symmetries.

²⁴³ If a meaning differs between two individuals (i.e., agents), then coordination between them will be impaired, as they will not agree on the information in messages received. Thus, coordination involves determining what messages mean, or at least assuming they will mean the same things to all parties. Thus, even though I state that the receiver must determine meaning for symbols, the transmitter of a message will often strongly influence this for receivers. However, note that this also requires the receiver understanding the transmitter's orders, which are themselves symbols. Regardless, the point is not to be taken as the receiver having a free choice to decide but simply that any receiver of a message must form a symbol set from which to process Shannon Information for the message.

²⁴⁴ Note: This chapter primarily discusses human or AI/ML transmitters and receivers, which are complex information processing systems. It is difficult to map brains or ANNs physically to this diagram. For example, one can loosely equate the brain to the role of receiver but could also say that the brain includes as many channels as there are axodendritic connections between neurons. This does not mean there is anything missing in the Shannon Information diagram. Similarly, the claims made in this chapter for Semantics as forming a macrostate may be an appropriate model but should not be taken to be something that can be pinpointed in the brain as a formal set of steps carried out mechanistically.

Figure 3. Schematic diagram of a general communication system. Figure taken from (Shannon 1948). The Transmitter, Receiver, and Message concepts are used throughout this chapter. Note that, in Shannon's model, the Receiver goes to the Destination and the Information Source goes to the Transmitter. In what follows, I will ignore the Destination and Source for simplification purposes.

Likewise, consider a game of 'Twenty Questions'.²⁴⁵ If the answer is 'Aunt Gladys', then a question revealing the person's gender provide less Shannon Information than the person's first name. However, this is only true if both gender and names have some symbolic meaning within the guesser's macrostate. Abstractly speaking, we can state that the model that comprises a guesser's macrostate for an environment (i.e., a particular game of "Twenty Questions") requires an **index** of symbols to gain Shannon Information. Without the index, there is no way to compute how much information is gained upon receipt of symbols. In fact, what is received cannot even be called symbols from the perspective of the receiver without indices against which to map them. We will see that this is critical to defining Symbol Grounding. We will also see that understanding intrinsic information processing requires considering how the models get defined in the first place, or what I refer to as "forming" micro and macrostates. My stance is that Information cannot be measured without an observer. Even if the entropy was the same for any observer given the same system modeled, the observers are needed to define the model.

Shannon Information is Non-Semantic

Shannon Information has not only been critical to the development of computers and communication but is also leveraged in AI models, including for measuring uncertainty and optimizing prediction and inference. Unlike most uses of Information Theory, for AI²⁴⁶ we are often interested in measuring Shannon Information from the perspective of the AI model. Certainly, at least for AGI, a criterion for intelligence is to extract information from observations and the ability to account for uncertainty for inferences or predictions about the world.

With this in mind, we must first note that the equation for Shannon Entropy is non-semantic. It says nothing about what the messages measured mean, only their relative probability of occurrence or the probability of the symbols representing possible microstates given the symbols

²⁴⁵ 'Twenty Questions' is played with two more people, where one person has an answer in mind (e.g., a type of animal, such as a racoon, a family member, etc.) and the other person has to guess what it is by asking questions that the person with the answer can respond to with 'yes' or 'no'. The guesser wins if they determine the answer using twenty questions or less.

²⁴⁶ For the remainder of the manuscript, I will shorten 'AI/ML' to 'AI'.

received from messages. This fact was already established at the birth of Shannon Information Theory. Despite its fundamental and generalizable nature, Shannon and Weaver (Weaver 1953) realized it did not cover all properties one would intuitively apply to Information. They postulated at least three levels of Information in need of theories: Technical, Semantic and Effectiveness problems. Shannon's Information Theory only covered the technical type, which I have and will continue to call "Shannon Information".

Briefly, the other two refer to: What symbols mean (i.e., Semantics or what some have called, "Semantic Information") and the significance or value of information, which I refer to as "effectiveness" to associate it with effectiveness as it is used earlier in this manuscript. Of these types, considerable work has been aimed at defining Semantic Information, either on its own terms or as derived from Shannon Information. I will describe these later in this chapter. For now, I make no assumptions about whether Semantics should be a different type of "Information". Since I do not define a more general concept for Information, I will only refer to the Shannon (i.e., Technical) type as "Information"; and the other two will be described as 'Semantic Processes' and 'Effective Processes', unless otherwise stated.

However, I can claim at this stage that Semantics is distinct from Shannon Information. For the reasons described in the previous section, **without semantics, there are no symbols to ground**. Specifically, for messages received to be measured using Shannon Information, symbols must first be determined for the elements contained in the messages. Shannon Information is fully conditioned on what those elements are determined to represent (i.e., what they mean). We saw this earlier where the elements in a message could be numeric values for the number of dots or the number of rotational symmetries for the faces of die. There are an arbitrary number of possible semantics for which to determine what is symbolized by elements. Indeed, this is why I state that it makes no sense to call the elements 'symbols' until they actually symbolize something. Furthermore, there is no clear way to emerge semantic properties using Shannon Information Theory. The same is true for Algorithmic Information Theories (Hutter 2007), which also operate on symbols at their own level, never the content they evoke. Any semantics only come along for the ride through the mind of external observers. The result is a lack of a theory for how this necessary process can be implemented in AI.

This creates an issue for AI and Grounding. For the most part, the issue has been ignored or confused due to the natural inclination to project our own semantic context onto the systems we develop, just as we do with most computer programs. And yet, the ability to measure Information from the reference frame of the model will be critical for AI, *including the different emergent levels they must conceive*.²⁴⁷ In a later section, I discuss how Semantic and Effective Processes can be described to resolve this challenge. However, I first need to reemphasize this issue of projecting emergent (semantic) properties onto external systems, so it is kept in mind for what follows.

²⁴⁷ Research: (Gleick 2011) claims there is a paradox in information: As we gain more information about the Universe it seems to have less meaning. Consider this in the context of the issue of Shannon Information being divorced from Semantic Meaning, as well as how Information relates to Eliminative Materialism.

Information is Emergent

As already stated, from an evolutionary perspective, it is effective to perceive the properties we project onto external stimuli as properties of those stimuli.²⁴⁸ Emergent properties are heuristic representations that help us cope with these systems. However, for many types of external stimuli, this can lead to the illusion that those emergent properties are emergent intrinsically to them. This can create considerable confusion, particularly in developing AI, where we build models with the objective of *them* processing information and understanding semantics.

Information is emergent and requires an observer. Since Information measures a decrease in uncertainty, there must be something that is uncertain, which only makes sense if some thing is uncertain about something.²⁴⁹ As said by eminent physicist, Donald MacKay (c.f., (Floridi 2011)), information must be a distinction that makes a difference.²⁵⁰ Something can only be a distinction to something intrinsically able to ground it as such. Information must be a distinction that makes some sort of measurable difference in the receiver's model of knowledge. However, this distinction cannot be assumed based naively on an external observer's extrinsic perception of that receiver's model. Information only makes sense with respect to (intrinsic) observation. "Observation" here can refer to external stimuli (i.e., environment) or internal stimuli (i.e., propagated thoughts) received, but there must be something that observes either way.

To make sense of this requirement for an observer for Information, we need a way to define what "difference" entails. Later in this chapter, I will propose what I believe can meet the criteria for there to be something that counts as a difference from the reference frame of a system itself.²⁵¹ However, I first want to describe what I believe are the implications for claims that anything processes information without having new criteria for grounding (e.g., with only having Shannon Information Theory to understand information processing).

To start, I will claim that many of the systems upon which we confer information processing properties cannot be said to do so intrinsically from our current understanding of Information. For many others, it is not directly clear. To take as an example, in Cellular Biology, molecular interactions are often described as information processing. When a cell receptor binds to a specific molecule, we often describe that molecule as a signal. However, since cells can be reduced to chemistry, why aren't all chemical processes considered information processes? One might assign special criteria to the living and its status with respect to evolution as consigning agency or even a purposeful quality. However, chemical processes have histories too. One could claim that the history for noncellular chemistry does not matter for explaining the properties of

²⁴⁸ Note: I will call these stimuli 'objects' or 'systems' to make subsequent claims easier to follow, but note that these are technically ambiguous terms, since the concept of an object and a system is already an emergent heuristic.

²⁴⁹ *Research: Nothing at the mechanism level possesses the property of uncertainty intrinsically. Uncertainty seems to be an emergent concept itself. How does it form?*

²⁵⁰ This is true for anything that would be defined as Information.

²⁵¹ As always, this will be with respect to grounding only. I make no claims about what is required to intrinsically experience stimuli, including as information.

the given system. However, “matters” implies value or at least predictability, which in turn is relative to the regularities of interest to an observer.

For biological cells, and for many other systems, the way we typically confer information processing on them by projecting our own emergent properties onto the systems. It is natural to describe cells, as biological products of evolution, as having signals and noise, but how we do so provides no evidence that these exist intrinsically for them. Instead, we need to establish appropriate criteria for what entails having an observer to process information.²⁵²

Can we really state that cells process information, and if not, does it make sense to refer to biochemical dynamics in and across cells as “cellular signaling”? **A signal assumes a preferred (or at least intrinsically distinguishable) state. That preference is intuitively related to a cell’s “function”, which, if pressed, may be described in terms of evolutionary adaptations. This leads to conferring teleological properties to the cell. Of course, this might be acknowledged as merely an explanatory heuristic. The problem is that this acknowledgement is not extended to the projection onto a cell to perceive it as receiving and processing information intrinsically, even though this is equally an explanatory heuristic, at least in how it is assessed. The evolutionary process does not necessitate that cells *intrinsically* distinguish or prefer certain states over others. At the mechanistic level, the cells simply react differently to different molecules in certain, often complex, ways. This “difference”, however, does not necessarily require intrinsic information processing by any of molecules or networks of molecules.** If there is no intrinsic processing, then calling a cell a “signal processor” is like calling ice a signal processed by a drink to make it colder.

Referencing energy minimization does not resolve this issue, as it would first require explaining how the cell needs to be assumed to *process* each molecule type with respect to a model to update entropy. This process can occur completely naively without the cell having a model, and the description of energy minimization is only understood by the external observer. **A cell’s chemical makeup does not necessarily reduce uncertainty for the cell.** We cannot use just our own extrinsic emergent projections, from our own reference frame, to determine if information is being gained by the cell as an observer (i.e., intrinsically). Without intrinsic experience, the cell also cannot have its own extrinsic emergent properties. **Basing any external system’s processing capability directly on our own heuristic projections fails.** From such criteria, terms like “messengers” and “information processing” are anthropomorphic. A signal entails a receiver that places special status on particular stimuli. In the case of a cell receptor binding to specific molecules, we project a goal of binding that molecule, which converts it into a signal, from which we infer the emergent concept of information processing.

If a cell does not intrinsically discriminate signals, then it does not intrinsically perform information processing. Information requires an observer (i.e., receiver), but it alone provides no criteria for that observer to be said to itself process the information it receives, and thus to count as information intrinsically. The same must be said about neurons and even neural networks, even though they are supervened upon by the observer’s mental concept of

²⁵² I will propose criteria for this later in this chapter.

information (the one with the goals). This provides insight into the problem of Mind: Consciousness, and many of the concepts we use with it, are more observer-centric than we realize, to the extent that it includes the notion of Information itself. **This further suggests that we cannot apply the term Information Processing to computers using our default extrinsic emergent projections onto them as criteria. Furthermore, there must be something else to provide the criteria for counting as Information beyond Information itself.**

What this all means is that, *given our current theories on Information*, the external systems we claim process Information cannot be said to really be (formally or informally) measuring their own Information processing. Of course, such an ability is what we need for AGI. **All measures of information processing I am aware are external measures; external observers (semantically) selecting a model against which to measure the externally observed neural dynamics.** To do so requires an observer, and nothing within our models of Information set forth non-circular criteria for something to count as an observer (i.e., to model a distinction within “messages received”). Consider that, if Shannon Information did not require an observer, then entropy measures would be undefined. The theoretical set of animals behind my door right now is infinite in theory (i.e., across possible worlds). Something must determine the model for which information would be measured.²⁵³

There is another perspective from which we can see how information is observer-dependent. The state spaces of many complex systems have information values that exceed well beyond the number of atoms in the universe.²⁵⁴ If the information belonged intrinsically to the agent coping with the state space, then this would be a problem. In such cases, one could not believe that the agents are computing those specific information values intrinsic to themselves, even if they were able to measure information. Astronomical information values are possible for us to measure conceptually for external systems because many of those external systems, including the Universe itself, have dynamics that do not require a model of the environment. Any model, such as the type required to process information, has limited capacity. Thus, we can compute how much information would be processed for a system as a model externally, just as we can measure astronomically low probabilities for events. Such measures are external constructs about those systems; thus, they experience none of the physical limitations to exist “out there”. For example, Darwinian Evolution does not have to compute itself, which makes it freer to generate incredibly complex processes (e.g., brains).²⁵⁵ Going back to the Energy Minimization example, many of the

²⁵³ As we will see, this involves semantics to form the model.

²⁵⁴ Seth Lloyd claims that the Universe consists of 10^{92} bits of Shannon Information (Lloyd 2006). More generally, we can add possible events to systems (e.g., an entire human quantum tunneling through a wall to another room) for which the information gain could be higher than the information capacity of the human or even the entirety of the Universe. The point is that what we would measure as a theoretical model for such information processing clearly cannot be the same as what the human models to measure information if it measures information at all. So far, there is no reason to believe our AI models are different, even if did have the requisite capacity we are assuming.

²⁵⁵ The claim that Nature does not need to compute itself and so does not have this overhead does not mean that it can be arbitrarily complex. For example, it must obey the Laws of Nature. Also, if there are too many interactions, then there is the potential for instability. This means that many natural systems will be sparsely connected. However, these constraints have nothing to do with Nature actually computing itself.

systems that minimize energy in Nature cannot be said to be actually be measuring how to minimize energy (or how much they do so), even though we can compute it for them externally. If they did, then it would often be intractable. Instead, we as observers can attempt to measure it. **It is only when we model those systems that information comes into the picture, which is why models require approximations.** Just as this is true for Nature, it is true for any external system. The point I am making here is that this claim for energy minimization goes for information processing more generally. **If all the systems we conceptualize were intrinsically processing the information we measure for them externally, then many of them could not exist.**^{256,257}

The role of the observer is also clear in determining how phenomena are grouped. There is no sense in which information is measurable for objects not grouped into some membership. However, to group objects or events requires biases towards what properties to consider (and at what resolution). Consider communication, where bit sequences representing letters can be used to emerge words in the mind of the receiver. The amount of information measured at the

²⁵⁶ Note that I do not address quantum uncertainty in this manuscript, including if it is observer-dependent. However, quantum information must address the same issue. Paul Davies (Davies 2014) provides a thought experiment where the information required to determine the probability of a photon going through a series of beam splitters requires more information than is available in the universe. One can set up 500 beam splitters for a photon to go through but cannot model the full state space because it would require more than all of the bits theoretically available in the universe. This demonstrates that setting up the experiment is clearly not the same as the model, since only the former in this case can be done with a feasible level of resources. Since the experiment itself can be done in theory, it must not require computing its own information (i.e., modeling itself) to implement. Only models can bring information to bear on a problem, and in turn are limited by processing power. If this argument is valid, then it extends to many other areas. For example, irrational numbers are impossible to fully model. How does the ratio between a circle's circumference and diameter attain such a value? The answer must be that computing π requires an observer to measure it or to form a model measuring it. The processing limits apply to a computer calculating π , which is clearly not the same as a physical or Platonic circle having π . The former is a measure, not something that exists "out there", or at least not as something that must be intrinsically computed for the property measured to exist. Calculations and models are tools that observers apply to the world, and emergent properties count in this category of observer-dependent models. What about infinitesimals? Can these exist without requiring the Universe to intrinsically perform associated calculations? Does separating our conceptualizations of phenomena from phenomena themselves help resolve any outstanding ontological issues?

²⁵⁷ As an aside, although it is unclear (and maybe impossible), there is research suggesting that biological brains can be Super Turing Machines. (Siegelmann 2003) claims their concept of an Analog Recurrent Neural Network is not only Super Turing but offers the maximum computational power for Super Turing machines. The concept includes three criteria that must be met to be Super Turing, but they all seem to point to the fact things in the physical world are their own best "model" (by which I mean no model could fully capture it), specifically because their processes are defined by real numbers (rather than natural or rational ones), e.g., for their neural weights. In other words, they are physical and so consist of real-numbered weights, which can impact function in ways that digital models cannot. We could presumably never digitally model them or even precisely emulate them in analog since we do not have infinite precision in our measurements of these weights. Again, this points to the territory not needing to map itself. However, it also suggests that there is at least the potential for considerably more computational power if we: 1. Build physical neural networks to drive optimization; and 2. Leverage their analog properties, so that the potentially infinite precision that is used in the physical system is leveraged for those objectives. This also relates to Gödel's Incompleteness Theorem (Gödel 1934), which Gödel took to mean that mathematics involves something that is not captured with logic alone. Turing Machines are built on logical instructions, and so it may be that certain non-Turing designs could in theory capture something more. However, the more precise the analog values must be to enable what we would want them to do, the more improbable it is that we would happen upon values within the needed ranges. It is precisely because we cannot model them to this level that they have Super Turing potential.

emergent level is not contained in the lower-level mechanisms alone. The receiver is required to determine information gain.²⁵⁸

Once again, this issue arises when conflating third-person with first-person views. Envisioning a person who is not observing certain “signals” already envisions them as emergent “signals” with statistical properties. In other words, Shannon Information from the third-person view assumes a receiver, even in this case where the explicitly referenced receiver differs from the implicitly assumed receiver. **More generally, Information and Entropy are properties of a conceptualized system, but the entire concept of a system requires an observer, and the information and entropy properties are projected onto this emergent system as extrinsic emergent properties.** We can now relate this to the claim that Statistics is emergent. Probability involves statistics, and statistics only make sense for systems. Any statistical properties are emergent phenomena measured with respect to the observer's uncertainty about that system. This typically is associated with macrostates and depends on it and the chosen microstate properties. Likewise, **Thermodynamic Entropy can be considered a measure of the degrees of freedom for a system. However, the physically realized microstate does not have any true degrees of freedom (assuming classical mechanics). It is only at the macroscopic level that degrees of freedom are meaningful.**²⁵⁹ **The macrostate is always a property that a human can project on those systems (i.e., an extrinsic emergent phenomenon contained within a human's intrinsic emergent experience).**²⁶⁰

²⁵⁸ I add one more example to drive the point home: Discarding extreme high and low frequencies in audio and video recordings reduces the amount of data but does not affect quality for the observer who cannot hear them. Yet, in terms of Shannon Information, this compression is still lossy, depending on the model applied. That is, for a model that includes those frequencies, Information is still lost when these frequencies filtered are not transmitted all the way to the receiver. In this case, one might claim that there is no information loss for the observer since anything unobservable is not informative. However, choosing to ignore these variables or variable values is a modeling choice that relies on a receiver to not only process information but to determine what information can be processed. Some observer is needed to determine what variables to include to measure Shannon Information, but we see here a clear case where the observer can be something external to the external system of interest. Shannon Entropy and Information Gain can be arbitrarily determined for an external system by changing the range or granularity of the model variables, and it can be measured and projected onto the system, whether or not that external system processes Information itself.

²⁵⁹ More accurately, one can of course form a microstate that includes the property of uncertainty, but this is of a different conceptual form than the concept of degrees of freedom discussed here.

²⁶⁰ Claiming an observer is required for thermodynamic entropy to make sense is not the same as claiming that entropy is arbitrary. First, for all we know, there are a fixed set of properties that compose reality. If so, then those properties place non-observer dependent constraints on what an observer would model for microstates, at least to some degree. Secondly, I cannot think of a case where a closed well-mixed system ever reverses entropy over time, which should be possible if any model could be constructed for entropy. Symbols formed by an observer for a physical system can change the initial entropy, but any such symbols seem to obey the Second Law of Thermodynamics. I have not comprehensively considered counterexamples, but my current belief is that physical mechanisms constrain formable micro and macrostates, and the observer cannot choose how entropy changes for those systems once the index is formed. Shannon Entropy may be able to be chosen more arbitrarily. This might be possible by not having the closed system requirement of underlying physical processes. For example, if I have one messy and one tidy roommate, I could claim the distribution of laundry on the floor is not well-mixed. This is a valid Shannon Model, but the physical properties involved here are an open system, given entropy expended by the energy put forth by the tidy roommate cleaning (among other physical variables). The difference seems to be that

To summarize, without determining criteria outside of Information Theory itself for possessing the ability to observe, there is no evidence that biological cells, AI, or any other external system, have a capacity to intrinsically model anything, which is required to have the capacity to measure uncertainty and process information from their own reference frame. Our belief to the contrary is due to our projection of emergent processes onto those systems. In such cases, *we, not the systems themselves*, are modeling the information processing of those systems. Only if it could have its own point of view could we say there is the potential for uncertainty, where information influences models formed (again, intrinsically).

All we know directly for external systems is that they have the capacity for bearing information for an observer.²⁶¹ And yet, the universe made information processing agents; and all empirical evidence points to certain biological cells as performing that information processing for them. How can we reconcile lacking evidence of an observer in these systems with the fact that we, as physical agents, perform information processing? All the above claims could suggest that Information Processing requires a first-person view. If so, then we are not equipped to assess it for external systems for the same reason we cannot assess intrinsic emergence for them. However, I propose below an alternative approach that sets criteria for intrinsic information processing and symbol grounding. What matters for the current section is that **the qualification of whether something intrinsically experiences information should not be based directly on what emergent properties we confer on them. Unless we have other criteria for intrinsic information processing, to describe cells or AI as information processing systems is a heuristic representation that we project onto them.**

An Alternative Approach

If information requires an observer, and we cannot make intrinsic claims from the third-person, how can we ever measure information in external systems? What we need is a way to measure from our (i.e., human observer) external point-of-view how a system (e.g., AI model) grounds concepts, but one that does not naïvely rely on the extrinsic emergent experience. One possibility is to make no claim for an object having phenomenal experiences but instead what its grounding would be if it did. In other words, we measure the properties from our own extrinsic viewpoint without making any claim about them being a part of (or accompanied by) phenomenal properties. For example, whether there is anything it is like for a cell to experience properties of its environment is unknowable. However, one can ask if cells ground their environmental

thermodynamic entropy must be constrained by the underlying physical processes to account for what counts as a closed system, whereas Shannon entropy, while constrained by the physical processes, just as any emergent phenomena would be, is not constrained to directly *account* for the actual physical processes in such a way that a closed system can be determined. This places higher stringency on thermodynamic entropy, while Shannon entropy need not bear any obvious relation to underlying physical microstates (e.g., a Shannon Entropy model for sentiment expressed in social media posts, the sum values of two 10,014-sided dice, etc.). Regardless, this has no obvious bearing on the claim that all forms of entropy require an observer. A state is only defined with respect to some choice to go beyond the individual elements and “decide” to group them in some manner. Requiring an observer does not mean that what the observer envisions is not constrained by the individual physical processes that underlie the state, however indirectly.

²⁶¹ *Research: Consider if/how prediction is itself an emergent concept adapted for its effective value.*

interactions. If so, then I argue that it is appropriate to suggest that cells process information. Specifically, if a system meets the criteria I will propose, namely that they perform Semantic Processes, then they can be said to intrinsically process information. This differs from whether or not a system is experienced by us as processing information.²⁶² The way information and signal processing are usually conferred upon systems has more to do with the extrinsic emergence we project on them. Some of these may meet the criteria I set forth to be considered to ground symbols and therefore be able to process information (i.e., be observers) and some may not.

Again, **the solution to associate grounding with information for an external system/model is to measure the external system as if it had the necessary agency for intrinsic information and thus a model with the capacity for information processing. However, how we assume this requires meeting certain criteria, which do not directly relate to the emergence we project onto the system.** With this in mind, I will simplify this tortuous logic by assuming the reader will keep in mind that anytime I mention grounding or emergence for models, I am referring to how we extrinsically emerge the given model as having this property of grounding and emergence. I am making claims about what is required to have intrinsic processes but not whether there is any phenomenological experience associated with them. The grounding process can be defined externally, as we will see.²⁶³

Criteria for symbol grounding are used to assess whether external systems can intrinsically process information. The reason is that grounding is required to determine the symbols that form macrostates and microstates, which are necessary and sufficient for information processing. Thus, when we can claim that external models ground stimuli, then we can claim they have the capacity for processing information intrinsically.²⁶⁴

The criteria I pose for a system to be said to inherently process information is that it must be able to (explicitly or implicitly) form a model to serve as the macrostate against uncertainty is possible. Forming a model requires determining the indices against which the model is updated for uncertainty to change. Further, when we, as observers, predetermine the indices for such systems, there is no longer an *intrinsic* role for information processing for the system. The ability to externally perceive such systems as processing information is possible, but it makes no sense to say the system itself is doing the processing. Something external to the system determining the symbols converts the system to a model, not a modeler.

²⁶² Although the potential for a relationship between the two criteria is worth considering.

²⁶³ Note that, while I made arguments in Chapter 1 against any digital system having intrinsic experience, I also stated that it cannot rule out some “consciousness ether”. While I do not make any claim here about this ether, we can use it to distinguish the conservative hypothetical I am making above vs. Explicit Representational Models (ERMs): If there were some “consciousness ether”, I could not rule out models as having intrinsic grounding proportional to the information-related measures I establish in this chapter, but I can essentially rule out them having grounding when populated with explicit representations (i.e., labels). By “essentially rule out”, I mean more specifically *that the grounding we project on these models cannot be intrinsic to the model*. ERMs would still have zero or low intrinsic information, as we will see. The information we ground those labels with extrinsically is not available to those models, but it might be for other model types, particularly the ones we might want to develop that align with the AGI requirements claimed in this manuscript. This justifies using this hypothetical to move forward with the measures discussed in this chapter.

²⁶⁴ Research: Consider further the similarity with quantum phenomena with the respect to the role of an observer.

If I am correct that we cannot predetermine the indices for a system's models against which messages become symbols relative to those models, then this means that a message from the environment cannot itself intrinsically bear objective symbols. The messages are only symbols in relation to the formed models and thus can change depending on the system that processes them (i.e., the receiver). **Messages can constrain what models are formed, but they are not symbols until a model is evoked. I claim this process of constructing models against which messages achieve symbolic meaning is the process we should identify with grounding.** This entails that, for grounding to occur, the messages must be subsymbolic. However, subsymbolic here is also relative to the models that will symbolize them. Thus, we can propose that, for a system to process information and be grounded, the messages received cannot *a priori* bear the symbols with which they will update the system's macrostate(s). The system of interest must form the models evoked as well as the symbols they apply to the messages received or there is no grounding or intrinsic information processing. **This criterion differentiates models from modelers and puts many AI types into the model-only category. It also means that it is the modeler that grounds symbols, not the model.**

This need for the agent to receive messages that are outside the symbol set (indices) that will be used to form the models (i.e., macrostates) may seem arbitrary. I believe the reason is that we naturally assess it from the outside, as discussed for Extrinsic Emergence in Chapter 1. Any phenomena could be incorporated into a single model; a model could include 0s and 1s to capture messages, say, as raw bit sequences, and this same model could include concepts such as velocity. However, I believe the issue here is that we, as human onlookers of the agent (including thinking about the agent while reading the claims made here), are assessing the agent, from its processing of messages to building models, from the third-person, where everything can be considered symbols (as grounded by us), which are therefore concepts and eliminable. The only way I can see to begin to make sense of what I am claiming here is for readers to each consider themselves as this agent, where everything involved in the message processing and model building processes occurs from the first-person view. However, I will put forward a more complete process to assess the criteria for intrinsic information processing through grounding later in this chapter. We will see that this requires defining Semantic Processes as pre-information phenomena.

Semantics/Grounding and Semantic Information

Overview of Claims

This is the longest section in this manuscript, so I will start by summarizing the major claims:

- Semantics/Grounding requires that the AI (i.e., system of interest) receives subsymbolic stimuli from distinctions (data/elements in messages) to form microstates and associated macrostates (models, which is an act of interpretation). The formed macrostates can be modeled symbolically and thus their updates can be measured using Shannon-like Information.
- For processes to be defined as 'subsymbolic', they can be symbols from the perspective of others but not for the AI in question. To be so requires that the dynamics that

determine micro and macrostates are not themselves symbols in the micro and macrostate indices.

- Like Shannon Information, existing Semantic Information Theories start at a conceptual level divorced from the fundamentally perceptual Semantic Process occurring internal to the receiver's model. Starting with symbols converts processes into concepts and *assumes* Semantics rather than explaining their role. This has functional implications, corroborated by the inability for symbolic models (or use of symbolic information theories) to meaningfully relate symbols based on semantic properties or to learn new grounding for symbols to apply against data.
- To reach this conclusion requires addressing the issue that any digital model can be described end-to-end at the level of bits. I do not attempt to address whether intrinsic emergence can be experienced by these models. Instead, I focus on whether there is a functional analog to such experiences that could be measured by an external observer (i.e., a human) for the system of interest (i.e., AI model). If a system of interest meets the criteria I propose, then they can be said to both ground symbols (i.e., possess Semantic Processes) and process information intrinsically. A system can be categorized as a Modeler only if it meets these criteria, as opposed to simply being a Model.
- While these criteria are inevitably extrinsic emergent assessments, the external observer is not assessing grounding based on how semantic the outputs and their associated "processing" seem to them as observers. Instead, the criteria require taking the theoretical reference frame of some system of interest itself and determining the relation between symbols and the process by which they are formed. A Conceptual Space or Semantic Vector Space formed by an underlying Neural Network would be found to meet these criteria, while ERMs would not.²⁶⁵ For ERMs, what symbols map to messages are already determined external to the AI model, which means the model itself provides no additional model forming (i.e., interpretation) given what is received. In contrast, Conceptual / Semantic Vector Spaces must determine what symbols to map to messages. That is, they determine the micro and the macrostate(s) to then update via Shannon-like Information.

Due to the inherent difficulty in explaining these claims, some are repeated often as additional points are made to which they relate.

Preface

Clearly, Semantics has a critical role in claims made in the chapter. AGI needs to understand what messages they receive mean. However, current Semantic Information theories do not capture what is critical, which is Grounding. They also measure Semantic Information against some objective truth or intention that is external to the mind of the recipient for which we want to measure it. That is, they are counterfactual measures. This removes the measure from the

²⁶⁵ Conceptual Spaces and Semantic Vector Spaces were introduced in Chapter 3.

very model against which we want a measure. They also cannot escape the problem of the emergence being projected onto the models by the humans. However, Shannon Information itself is measured in this way (i.e., Shannon Entropy is measured for a probability distribution over the set of possible symbols). I am not about to suggest redefining Shannon Information, and it would be confusing to keep Shannon Information as a counterfactual measure but make Semantic Information and Effective Information non-counterfactual. Moreover, none of the claims in this manuscript contradict Information as a counterfactual measure. Ultimately, it seemed more appropriate to not describe Semantic and Effective measures as “Information” and to simply refer to them as “Processes”. In doing so, however, I deviate from Shannon and Weaver’s claim that these are other types of Information (Weaver 1953). It may not matter if these processes are referred to as types of Information. After all, these are just terms. However, since Effective Processes shape Semantic Processes, which determine Shannon Information micro and macrostates, it makes sense to think of these processes as pre-Information measures. Information theories all associate Information with symbols, so I avoid using the term ‘Information’ to describe Semantics.²⁶⁶

I therefore reserve the term ‘Grounding’ or ‘Semantic Processes’ (or just ‘Semantics’ for short) for a noncounterfactual form of Semantics. ‘Semantic Specificity’ is a quantifiable measure (a magnitude) I will also describe for Grounding. Grounding will be associated directly with Semantic Processes rather than Semantic Information. **I likewise reserve the term ‘Effective Processes’ for processes that shape Semantics and Shannon Information based on effectiveness for the receiver (or the receiver’s ancestors).** This also allows us to attempt to relate these concepts to existing concepts of Semantic Information, rather than replace them. However, I want to note here that what I think fundamentally matters for Semantics are the processes described in this chapter, which seem to be missed by existing Semantic Information theories.²⁶⁷ As long as this section proposes a way forward to measure grounding, and as long as that measure has the role I believe it must have for the receiver, it is not important to associate the term “information” with it. What is important to convey, however, is that, whether we call it

²⁶⁶ This does create one complication. What existing Semantic Information Theories refer to as ‘semantic’ is the level of meanings associated with facts of the world (e.g., what a dog is, whether a television is broken, etc.). While this is the richest level of semantics for humans, I will propose that it is only one level of semantics. Semantics is more general. It determines symbol indices for any Shannon-like model, which could be at the level of letters, words, facts, theories, metatheories, and so on. Thus, it would perhaps be better to replace the term ‘Semantic’ Information Theories with something specific to the emergent domains against which they are applied.

²⁶⁷ My focus in this chapter is on what it would mean for an AI model (or any model that is a modeler, such as humans) to process information intrinsically (“intrinsic” here meaning from its own perspective). It is possible, though never mentioned, that authors for information theories and associated concepts, such as Semantic Information, intentionally develop theories for how we, as third-person observers, should evaluate the processing of models as external models. This chapter is not an attack on such purposes. My claim is just that these theories for measuring external models/systems are not what we need to measure for intrinsic AI processing. We must be careful to not project our measures of them as measures they are making themselves. Doing so conflates extrinsic vs. intrinsic processes. As already discussed, having intrinsic processes makes a functional difference, including for generalizability, adaptation, and many other intelligent capabilities that non-modeler AI cannot achieve. We will see in this chapter why I claim that existing theories do not work for measuring intrinsic processes, such as Grounding.

Information or not, having a measure for Grounding, which is the real interpretation process, is what really matters. Semantic Information will not be defined as Grounding, but Grounding is critical for Semantic Information.

Put another way, something is missing with current Semantic Information definitions. My claim is that they ignore this more fundamental process of Grounding, and this limits their utility. If this is an oversight, then it may be due to the issue of necessarily describing grounding, a subsymbolic²⁶⁸ process, with symbols, which are needed for communication. This compounds with the issues described in earlier chapters of the inclination for humans to project our own grounding on our models, and theories behind those models, thereby possibly leading some to assume that grounding is accounted for in existing definitions. This creates considerable confusion, since we are always considering the symbolic level with our examples and cannot easily separate out our grounding that accompanies those observations.

It turns out to be tricky to separate any Semantic Information process from processes that could be described using pure Shannon Information. Below I present what I believe is the best shot at a theory that disentangles the knots. It does not make for easy reading. However, because Grounding is one of the critical needs for AGI (and for understanding ourselves), because it is missing in current Semantic Information theories, and because it is so difficult to (necessarily) describe something subsymbolic like grounding using symbols, I dedicate a large amount of content to it.

Notes on Symbol Grounding

To distinguish Effective and Semantic Processes, I use ‘grounding’ to refer to semantics only. On the other hand, I assume that both Effective and Semantic Processes deal in **Affordances**. Affordances, which are properties in the environment that, often through interaction with agents, make information available, are more general than the way that I use the term Grounding. Unlike grounding, affordances are not always associated with experienced qualities (e.g., biases for how to act given environment cues) and cover the range of processes for how we use, and make available, stimuli to cope with the world. They therefore cover a lot of what Hubert Dreyfus and Martin Heidegger describe, as discussed in Chapter 2 (Dreyfus 2007). Grounding is associated with the experienced qualities provoked by signs (including symbols), while Effective Processes can constrain Grounding (i.e., Semantics).

I will use ‘Grounding’ interchangeably with ‘Semantics’ or ‘Semantic Processes’, since my stance is that Grounding is core to Semantics. Grounding, Context, Meaning, and Semantics are assumed to be equivalent. I have found no reason to separate them. Note that I avoid the term ‘Pragmatics’, which is associated with some aspects of Context in Linguistics I associate with

²⁶⁸ I use ‘subsymbolic’ to describe the nonsymbolic processes associated with grounding because they give rise to symbols (i.e., explicit concepts). However, the use of ‘sub’ should not be taken to imply that these processes are less foundational to cognition/intelligence than symbolic processes. Also, I have not attempted to separately define nonsymbols. Nonsymbols that are not subsymbols would presumably meet the criteria I set for subsymbols, which could be a subset of nonsymbols. Addressing what distinction there is between them has no obvious relevance for the claims in this manuscript.

Semantics but can also be taken to imply properties I associate with Effective Processes (i.e., the value achieved given information vs. context evoked/constructed for messages/data received).

Grounding is often associated with some mapping to real world processes. For example, “Grounding” in Robotics typically refers to mapping symbols to physical observations and actions on the world (Coradeschi, Loutfi and Wrede 2013). It might be assumed that symbols must map certain sensory experiences to categories that can then be demonstrated to achieve human-defined objectives. When this involves subsymbolic processes, such as ANNs to implicitly map sensory features to symbols, it can meet the grounding criteria I propose. **The ability to use any of our traditional set of human senses for the same types of messages as we would is likely necessary for AGI. Also, we would require AGI to be able to generally behave like us given those messages.**²⁶⁹ **However, I am describing a more general view on what grounding actually is that does not limit it to particular sets of senses**²⁷⁰ **(or motor actions) or how direct their assumed physical referents are. What specific grounding must occur for AGI is a different question.**²⁷¹

Luc Steels also seems to suggest that grounding only pertains to symbols about real-world phenomena:

“There are a lot of symbols which are not about the real world but about abstractions..., and so they will never be grounded through perceptual processes. ...I focus on groundable symbols.” - (Steels 2008)

What counts for Steels as “perceptual processes”? I take the view that there are *always* qualities underlying experiences of symbols, no matter how abstract. This includes a symbol of a car passing by, the abstract concept of corruption, the letter ‘A’ (I will describe why there is even grounding for alphanumeric units). Even some never-before-seen marking can be grounded as something I do not recognize. Given this view, either “real world” simply means anything that refers to something, or it seems to arbitrarily scope what counts to certain preconceived physical entities and events. However, the latter assumes the very concepts that grounding is required to conceive.

Even if we decided for some reason to come up with a consensus for what to count as real world references, it would run into issues very quickly. Assuming all symbols we experience as symbols were initially adapted or learned for referencing certain physical properties, those physical influences may be so ancient in evolutionary history that their current grounding no longer has any comprehensible link to obvious physical phenomena. Examples include

²⁶⁹ That is, within the range of human variance.

²⁷⁰ whether human or otherwise

²⁷¹ There is also the possibility of some researchers taking the real world requirement to mean that grounding is only necessary for low-level sensory/perceptual processing or that grounding is achieved for all downstream processes if these early layers are subsymbolic. As already mentioned, this has been common practice in Cognitive Architectures research. What are produced by these initial processing layers then become symbols used for explicit Reasoning. My claim is that, for processes to be grounded, there must remain a subsymbolic process that forms any symbols, including during any inference processes. Thus, this approach divorces the Reasoning itself from being grounded. I also claimed that grounding is functionally important for all cognitive processes, and critical for all intrinsic information processing, not just early-stage inference.

metaphors, such as someone “pushing” me to do something they want (Lakoff and Johnson 2003).²⁷² Stimulated with this phrase, my mind does not ground pushing with the physical analog of something physical pushing something else. This metaphor has deviated far beyond any initial resemblance in terms of the grounding it evokes.²⁷³ Generally, the effects of the external environment may be very indirect since certain grounds for stimuli may be conditioned on chains²⁷⁴ of other concepts.

Also, the grounding problem is not limited to, or solved by, the engineering of internal operations that are conditioned on receiving stimuli through specified sensing devices. This says nothing about how the agent is engineered to receive the given sensory inputs. Moreover, it is insufficient to assess whether grounding is achieved by relying on whether behaviors in the world met expectations of external humans (e.g., the model engineers). Note that a brute force model or simple symbolic rule could meet the criteria of achieving expected physical world outputs given predetermined sensory inputs. A brute force model seems more like avoiding grounding rather than modeling it. It would only work for simple problems, but it points to the larger issue of relying only on input and output criteria to define grounding. The only difference for more complex models would be that we cannot comprehend their process, but this says nothing about those processes intrinsically. I will propose that picking out the “correct”²⁷⁵ referent, while dependent on grounding, does not provide the direct criteria for grounding. It is insufficient because of the extrinsic emergence issues discussed earlier in this chapter, and it is unnecessary for a *general* theory for Grounding.

Thus, neither restricting the criteria for inputs nor externally evaluating the correctness of outputs are themselves appropriate criteria for grounding. It once again only measures grounding against the expectations of observers. Steels’ point may instead be just that a concept must be perceived to be grounded. As I argue elsewhere, any concept, no matter how abstract, is perceived as some instance, and this makes it perceptual. So, if Steels does not actually mean that grounding requires that perceptions be for certain specific phenomena, including what we might typically think of as more immediate sensory-based ones, then I would be in general agreement. Given how I use the term ‘abstract’, abstract concepts have perceived properties and therefore grounding.

Therefore, I assume the grounding problem is unresolved and take a more basic stance on what Grounding involves. The issues above are resolved by defining Grounding as Semantic Processes. Grounding becomes a useful process by compressing that which cannot be

²⁷² *Research: If all mental thought is governed by metaphors to external processes and properties (Lakoff and Johnson 2003), no matter how metaphorically distant the original referents are, then it is worth researching, when these are actually grounded in robots, how they can become internalized for mental processes (i.e., thinking). Grounding symbols for simulated planning by an AI agent may be a good intermediate step, since the planning could be about physical processes to execute but is still done internally. Then, move on to grounding processes that are abstract thoughts without any obvious direct relationship to sensorimotor actions except as metaphors.*

²⁷³ *Research: Consider how grounding concepts may be related to, and adapted from, how we learn to automate motor processes. Consider that grounding mathematics as symbols enables more efficient procedural use.*

²⁷⁴ A network is probably more accurate than a chain, but a chain emphasizes the dependencies more clearly.

²⁷⁵ “correct” typically refers to what the external human deems correct or what the human designed them to do.

comprehensively mapped by the agent/model, but its perceived usefulness or comprehensibility to us as external observers is not in the criteria. It involves the model determining how to decode messages occurring through application of implicit (and sometimes explicit) knowledge.

Semantic Information vs. Semantic Processes

While Shannon Information deals with the probability distribution of symbols, **Semantic Information Theories**²⁷⁶ attempt to deal with the probability of the meaning underlying the symbols. For example, the message “James is not a baron” has more Shannon Information (given a macrostate of English words as symbols) than “James is not a boy” because the symbol “baron” is less probable in this sequence than “boy”. However, the latter statement has more Semantic Information (at the level of a social domain²⁷⁷), since James is more likely to be the name of a boy, and therefore more greatly impacts the probable states of the world (i.e., who James is). If meaning is construed to be what symbols tell us about some domain, then this seems like an intuitive notion of Semantics. Related to this notion is that semantics involves logical rather than statistical probabilities. For example, “Socrates is mortal” provides information about the world that the occurrence of the symbol sequence “Socrates”, “is”, and “mortal”, does not. This meaning seems to relate to the message being treated as a logical assertion.²⁷⁸ A motivation for logical operations is incorporating truth values for assertions about the world. The semantic meaning for symbols is taken to be a fact about the world, and its semantic informativeness is associated with its logical probability.

My argument is that the core of semantics is not addressed with such measures, which is how symbols *obtain* their meaning. These measures only provide the probability given the meanings one would be expected to assign to them. That is, all Semantic Information Theories I am aware of differ in the level of microstates and macrostates but do not explain how the symbols and what they reference are determined. This makes it difficult to distinguish Semantic Information from a **Shannon-like Information**²⁷⁹ measure because there is always some level in which a set of symbols covers any facts of the world. A given level can be thought of as a **Reference Frame**²⁸⁰

²⁷⁶ Details are provided in a later section where I describe different Semantic Information Theories.

²⁷⁷ “social domain” is not a great description for the specific level I am trying to convey. I cannot think of a general term that conveys the world level of discourse for which assertions are typically considered to add knowledge about, and that we often think about as referring to what concepts “mean”.

²⁷⁸ I discuss specific theories with references later in this chapter.

²⁷⁹ I use this term ‘Shannon-like Information’ as shorthand throughout this chapter. What I mean by this is information as defined by theories that start with symbols. For example, existing Semantic Information Theories fundamentally model probabilities over symbols. They may differ in how they are formulated compared to Shannon or Algorithmic Information models. However, they are closer to these types of measures than the Semantic Processes I describe in this chapter that I consider a pre-Information measure. A clearer distinction would be simply Semantic vs. Information Processes, where the latter covers Shannon, Algorithm, Semantic, and Effective measures that start with micro and macrostates (and thus, with already determined symbols). However, I resisted this distinction to avoid readers focusing on what Information means given formal vs. folk uses of the term. The goal of this section is to propose a difference between the Information theories and true Semantic Processes.

²⁸⁰ Describing levels as reference frames will become important because it generalizes not only to what emergent level the symbols represent but also the difference between what we as observers symbolize for an external system vs. the external system itself. That it covers these two purposes that we would typically assume are distinct can be

from which one determines what are symbols (and to what they refer). Whether the elements in a set refer to letters, words, or objects in the world changes the entropy, but it does not alone add anything that cannot be measured using a Shannon-like measure. The microstate and macrostates (i.e., set of symbols and their entropy) differ for each reference frame. However, it is still a macrostate of symbols (i.e., a macrostate about the state of the world within some domain). Shannon-like Entropy can be measured in either case *once the appropriate micro and macrostate is formed*. I do not claim that Shannon and Semantic Information Theories are equivalent, but they all assume some symbol set from which to form micro and macrostates. It is this forming of the macrostate that is left unexplained by Semantic Information theories and will be the focus in how I will define **Semantic Processes (i.e., Grounding)**. Messages are distinctions (data) that evoke the *forming* of macrostates to determine what the distinctions symbolize and in what manner those symbols received update a macrostate as quantified by Shannon-like Information. The model forming process determines the impacts messages have on what macrostates. Both the macrostate and microstate must be formed prior to measuring Information. It is the formation²⁸¹ of the macrostate that involves semantics, and we will see that it is necessarily a subsymbolic process. A Semantics measure must address how the Shannon-like macrostates and microstates are formed in the first place. In summary, **the difference between Shannon and Semantic Information depends on the reference frame. What distinguishes Semantic Processes from Semantic Information is that the former forms which reference frame is applied.**

To be clear, the variety of Semantic Information Theories are useful, but they assume the core Semantics Process rather than explain it. There is a natural inclination to do so because, 1. We must describe any such processes using discrete and explicit symbols; and 2. We, as external observers with grounding capabilities, project the grounding that the symbols manipulated by the models evoke in us onto those models. When we as external systems determine the mapping for message elements²⁸² for a given model, then the model does not itself ground symbols and

confusing. The best way to think of it is simply as the theoretical perspective taken by a system/agent. Different perspectives can involve different levels or domains of emergent phenomena for which symbols refer. For example, “tiger” can be five individual letters, a single word, a single animal, etc. These are different emergent levels. However, extrinsic vs. intrinsic perspectives are also different reference frames. What is a percept of a tiger from an intrinsic reference frame can also be a concept of a tiger from a different reference frame.

²⁸¹ I use the term, “formation” to differentiate it from standard Model Selection methods in Machine Learning (ML) and Statistics. Model Selection methods do not necessarily meet the criteria I lay out for semantic model formation. In particular, if an external system is required to map symbol sets applied across the models and model selection process, then there are no subsymbolic processes and thus no grounding. This is evident from the need for a human to use their grounding to determine how messages map as evidence to update the models and model selection process. Aspects of some model selection processes may be grounded, but in general, the existing concept of Model Selection is not limited to processes that meet the Grounding criteria described in this chapter. In these cases, the model selection process itself can be viewed as a single model at a higher level, which includes the probability distribution over possible models. “Forming” is also consistent with the geometric forms in Conceptual Spaces that represent different semantic phenomena, as discussed later in this chapter.

²⁸² ‘message element’ refers to what in a message would be associated with an individual symbol. Note that this can differ by both transmitter and receiver. A receiver could also ground what are several elements into a single symbol. We cannot assume *a priori* that what we perceive as a message element is the same for some system of interest.

does not perform any semantic processing.²⁸³ This has a functional impact. **By ignoring the fundamental process of Grounding, Semantic Information measures have not really been useful for AI models. As in the Chinese Room Argument (Searle 1980), implementation of these theories does not address how interpretations arise, instead assuming something external from the AI itself to do the mapping for it.**²⁸⁴

By focusing on Semantics as a Grounding process, the following relations can be stated: Shannon Information assesses a set of symbols with respect to a symbol-level macrostate and microstate. Semantic Processes form²⁸⁵ the macrostate and microstate, which determines what the message received will symbolize. This is the process of *interpreting* a message, which sets the stage for the Shannon Information process just described.²⁸⁶ By forming certain micro and macrostates, Semantic Processes determine what symbols make up these macrostates and microstates, the updating of which given the message is a Shannon-like Information process.

Grounding involves implicitly mapping to what a receiver knows as context to determine what messages mean. Forming a model given a message is equivalent to interpretation. Interpretation relates to understanding/comprehension. Understanding can occur from different reference frames, such as understanding message elements as letters (e.g., 'd', 'o', 'g' = letters, words (e.g., d, o, g = word: dog), things in the world (e.g., d, o, g = animal: dog), and ever higher concepts (e.g., the transmitter believes they see a dog). Each choice of understanding can be thought of as forming micro and macrostates to update given the message received and is therefore equivalent to Semantic Processes / Symbol Grounding. Forming the micro and macrostates is also equivalent to choosing the symbol **index** for them.²⁸⁷ In the letters example above, the index could be the

²⁸³ **More accurately, the requirement to not explicitly predetermine (i.e., map) symbols for models is a consequence of the Description Gap rather than a criterion for Grounding. The result, however, is that any models that are modeled with predefined symbols are unlikely to also meet the criteria for grounding.** Graph Neural Networks (GNNs) are a possible exception in some cases where there are both explicit symbols predetermined and grounding through subsymbolic processes. What aspects of different GNNs meet the criteria is out of scope for this manuscript. Regardless, it is easier to understand the criteria I propose by simply saying that all symbols cannot be predetermined, since doing so is parasitic on the grounding of the external modeler.

²⁸⁴ In contrast, (Steels 2008) states: *"...in the Chinese room debate it is accepted that semantic processing can be done by computational systems whereas symbol grounding cannot."* I argue that semantic processing cannot truly be done by the system in question without grounding.

²⁸⁵ By model or (micro and microstate) 'formation', I do necessarily have some active construction or selection in mind. Model formation can simply be following some sort of physical gradient, where the gradient can be arbitrarily complex and/or stochastic. The obvious example would be a propagation of a signal across a neural network, constrained by synaptic weights. What is important to note here is that there is a process that determines what models are updated and therefore what information processing can be said to occur.

²⁸⁶ As a third process, **Effective Processes describe constraints that shape the forming (i.e., interpretation) process. Both evolutionary adaptations and experience will favor certain models to be formed through Semantic Processes given different distinctions in messages. In this way, Effective Processes, through their impacts on Semantic Processes, can also shape Shannon-like Information processes for the receiver.** Describing Effective Processes is left for a later section in this chapter.

²⁸⁷ Relating Semantics to determining indices for the Shannon Entropy equation is borrowed from (Haken and Portugali 2014), where they use it for explaining their theory for Semantic Information. While this helps grasp how Semantic Processes works, it could turn out to be more metaphorical than physical. The same is true for the idea of

letters of the English alphabet, where the microstate is a letter sequence, and the macrostate is then the Shannon Entropy over the possible letter sequences given the symbol set. The index could instead be the set of mammal types, where the microstate is the mammal type dog, and the macrostate is the distribution over all mammals.²⁸⁸ It could also be an index for the possible words the transmitter said (e.g., dog or dock) or what the transmitter meant (e.g., dog the animal vs. as slang for a friend).²⁸⁹

Grounding therefore relates to the process of forming models from knowledge. If I receive the message “dog”, and I have a model grounded with “dog” that relates to things I am allergic to, then this changes the set of possible messages from which to measure Shannon Information (e.g., grass, dust, etc., not the exclusive set of mammals). If something external chooses how data will map to symbols to update models, then the grounding only exists in the external modeler. With ERMs, humans predetermine the set of possible symbols. Shannon Information can still be measured externally for messages in these cases, even though the model does not know anything about messages received. It is the Semantic Processes of the human modeler that determine micro and macrostates and how they are updated given different possible symbols received, thereby avoiding the need for the ERM to *itself* be Semantic. However, this information is not the same as information that models might be processing intrinsically. The ERM is parasitic²⁹⁰ on the Semantic Processes of another system. Since existing Semantic Information theories also deal in explicit labels for concepts, they alone, like ERMs, require the humans to form the mappings from messages to labeled variables.^{291,292}

I take the general concept of **Information** as involving potential or actualized impacts of data on defined models. I believe this would be desired as a common base for any Information type. I therefore do not call Semantic Processes “information” because it fundamentally differs from Shannon-like Information. It is instead pre-information because it is required for a system to “symbolize” anything.²⁹³ Semantics deals in subsymbolic processes. To be semantic (and thus

micro and macrostates. That is, most cognitive processes may only implicitly construct and solve formal Information equations.

²⁸⁸ or those animals theoretically available to send by the transmitter since transmitters may have limited knowledge.

²⁸⁹ It should be self-evident that there is a relation here to Common Sense, which is associated with context one is expected use to model against a given message.

²⁹⁰ Use of the term ‘parasitic’ here is in some ways unfortunate. There are many cases where we have used ERMs to our own benefit. Overall, I believe it is more appropriate than not. There is a long history of investments in ERMs that may have been otherwise if this Semantics issue was clear. The survival of such methods for AGI and related objectives could thus be said to be facilitated by this human-dependence as a parasitic relationship. One could also rightly see it as a symbiotic relationship, but only in a very limited sense.

²⁹¹ This may be why Semantic Information theories have not been applied for any AI breakthroughs. They alone do not reduce the amount of human dependence.

²⁹² In Semiotic terms, both the signifier and the referent are explicitly defined by a human outside the models, and current Semantic Information Theories do not address how those determinations are made. My claim is that they miss the real semantic processes.

²⁹³ *Research: Consider how this pre-information role for Semantics aligns with theories that place a limited role for processing in perception (Gibson 1966) compared to standard assumptions about the brain and mind as information processors. I have not emphasized information directly available in the environment, and I have maintained an*

form micro and macrostates), there must be some subsymbolic process between the inputs received and micro and macrostate index formed. The meaning carried by messages is the subsymbolic grounding it evokes in the receiver. This grounding is associated with symbols from which micro and macrostates are composed.

Forming Microstates

It might seem odd that I am including microstates as part of the model formed through Semantic Processes. However, different possible microstates could have been chosen to measure the same data²⁹⁴. Micro and macrostates are both part of an overarching model that constrains the distinctions in data into the form of a state. **A State only makes sense within the confines of a model.** For example, distinctions between an 'A' and 'B' could be evidence of two different symbols (the letters A and B) or could be the same symbol (both examples of letters or both black-colored markings), or something else (something without curves and something with curves). These three examples have different impacts that depend on the model formed. Those impacts are the Shannon Information made available to the given models.

Each model has some possible set of symbols that could have been chosen. For example, even though common symbols, such as 'yes' mean the same thing across most models and may have the same symbol set ('yes', 'no', 'I don't know'), one still needs to determine the model

important role for Shannon-like Information, at least in theoretical terms. However, it is worth considering what issues with Information Processing are avoided or not with understanding of Semantic and Effective Processes.

²⁹⁴ Ideally, we would define Data before getting into how models are formed. However, I found this to be surprisingly difficult both in terms of satisfaction with existing definitions and in trying to define it myself. The challenge provides perspective on our own limitations. I would like to say something like, "Data is differences without their potential context actualized in the recipient observing it." Data becomes information as soon as it has a measurable impact to some system, often a receiver's model of some phenomena (that is, the distinctions in the data create distinctions in a model). This is consistent with Floridi's (Floridi 2019) description of Data but deviates in what exactly is required by the data to be informative. Floridi uses the term "well-formed" for this requirement, but this already assumes a model in the form of axioms. However, essentially any property for data seems to assume a model (e.g., context, differences, distinctions, nonuniformity). Data by itself should not be a relational construct. Relations only make sense in the context of a model. For example, nonuniform necessarily contrasts with a uniform model. What this all suggests is that **we, as modelers, cannot avoid understanding through modeling properties: We do not have the mental capability to think of anything without context. A definition of data necessarily invokes a model, but data itself should be model-less.** However, I think that we can conceive of data as a thing because we can imagine choosing different models against which to compare the same set of "distinctions". What is invariant across models is impossible to directly conceptualize but can be hypothesized as the unobserved that these different models actuate. **This pre-informative phenomenon, that we can only conceptualize in its informed state, is how we should think of data.** The advantage is that it leaves all the modeling properties to the models for which information will be actualized. **Data is that which is modeled and cannot itself be attributed modeling properties, including relations, regularities, etc. Data is what we are left with when the difference in the impacts across all possible models are excised from our perception of it, but to perceive is to model, so we cannot perceive data by itself, even if we know it is something.**

- **Research: Consider how the concepts above for describing Data aligns with the notion that Data could be described as anything that breaks some symmetry. For example, a perfectly homogeneous sphere with scratches in it, could be said to be data (Holland 1995). Symmetry breaking is associated with nonuniformities in the properties of the given environment. Does Symmetry assume too much as a definition for data (i.e., requires an observer model)?**

associated with the intention of the question to apply it to. In other cases, ‘yes’ could mean something entirely different (e.g., what is your favorite progressive rock band?). In all cases, the amount of information gain depends on the model formed. This again demonstrates that **Shannon Information cannot be measured until the micro and macrostates (i.e., model) are formed. Forming the model is the process of interpreting the message, which in turn makes the potential Shannon-like Information available through the chosen model for the receiver.**

The Role of Knowledge in Semantic Information vs. Semantic Processes

Some Semantic Information Theories rely on a receiver’s existing knowledge (Bao, et al. 2011). This makes sense given that Semantics provides context to determine how messages will be interpreted. However, such measures still say nothing about how an interpretation is chosen. They only tell us the outcome of choosing certain interpretations. **Interpretation is done through Grounding / Semantic Processes. In fact, Grounding is needed to measure *any* type of Information intrinsic to a system.**²⁹⁵ A highly specific message is useless (i.e., potential never actuated) if the receiver cannot map symbols as information for models. On the flip side, a message may be vague or incomplete, but the receiver may have knowledge that can enable a more specific interpretation. This can only happen if the symbols in and across the message can be grounded. Again, **Grounding is the critical measure because it is associated with the interpreting process, whereas Semantic Information Theories assume the interpretation.**

Interpretation as the process for grounding symbols means mapping them via some context within which they provide information (i.e., updating a model). If the symbols received cannot be grounded, then they are, in a sense, unknown to the receiver’s **Knowledge System**, and so a Shannon-like macrostate cannot be formed and updated given the message interpretation. “Knowledge System” here simply refers to what an agent knows, both subsymbolically and symbolically. Knowledge Systems are used by Semantic Processes to form models (i.e., micro and macrostates) that ground the message(s) received. Thus, this **Knowledge System is not merely a model of symbols and their explicit relationships.**

The specific grounding of symbols uses existing knowledge as a constraint for how messages are decoded. Less knowledge generally equates to fewer or less-rich²⁹⁶ models (i.e., macrostates) available to be evoked by messages. This is the same as the receiver’s Knowledge System. Facts and other distinctions are decoded using the Knowledge System actualized from messages and encoded as additional context for grounding future messages.²⁹⁷ As discussed later, symbols are what we are able to generate from the Knowledge System for inter- and intra-communication.

²⁹⁵ An external modeler can form models for a system, but this uses their own grounding and results in groundless AI with limited intelligence (since only explicit concepts can be modeled and not necessarily in the way they would be applied by a grounded AI that evokes them intrinsically. Moreover, only a limited number of explicit concepts can be modeled externally).

²⁹⁶ Technically, having less-rich models might be the same as having fewer models, in the sense that more rich models are unavailable.

²⁹⁷ **Research: Since Semantic Processes actualize potential (Shannon-like) Information, consider what a Knowledge System must have in terms of requisite levels (or perhaps more accurately, variety or complexity) to actuate the information available in messages. Otherwise, the model would “drop” potential information because it would**

Open-ended information processes are key for any intelligent agent to adapt in a complex, dynamic world. However, both Semantic and Shannon Information Theories consist of closed sets of symbols. Some other process is required for determining these symbol sets (e.g., a human using their own grounding to map symbols to model variables). For biological intelligence, such measures are not closed. Fortunately, Grounding is not addressed by a probability distribution of an explicit set of possible symbol sequences. A probability distribution can only be measured after the symbols for it are determined, which is what Shannon-like Information measures require. Grounding determines what the symbol set means (i.e., how they are interpreted / what they symbolize) by evoking the receiver's Knowledge System, which can learn new meanings for symbols that shape indices formed for future micro and macrostates. Symbols formed by the grounding can be used to update the explicit concepts in the Knowledge System as a macrostate, which I describe as a Shannon-like Information process.²⁹⁸ These symbolic updates can in turn change the grounding of the Knowledge System, which modifies the symbols that would be generated from this Knowledge System, such as symbolic assertions about the world to communicate to oneself or others.

Again, to be Semantic, the interpretation process must be done by the recipient's Knowledge System. If something external to an AI model determines how symbols, if received, will be mapped to variables to update them, then the interpretation has not been done by the receiver. Any such models should be measured as not grounding predefined messages.²⁹⁹

"Interpretation" here simply means forming the model to update. As I will discuss, Semantic Processes are not fundamentally conditioned on choosing some "correct" interpretation or even what the transmitter meant. Although these are common goals of interpretation, any grounding process is some interpretation. In contrast, **Semantic Information Theories typically assess how specific the interpretation was against some external reference (e.g., the set of possible interpretations). This makes them, like Shannon Information, counterfactual measures.**

A Counterfactual measure is an observer's assessment of some process external to it and therefore not of itself. If this is the AI modeling process we start with, then we are not modeling what we want to belong to the AI internally but instead some external model of the AI. Any grounding is undoubtedly shaped by many event types, but it is not a property that requires the

not know what to do with it in relation to existing knowledge. For the model to learn new things, it must have requisite variety or complexity to learn to map symbols, but not starting in the same form as the message symbols. This is exactly how ANNs learn, which do not start with explicit knowledge but have the requisite variety or complexity to learn how to ground symbols. Therefore, complexity and variety are not measured in the same symbols that encode messages, but instead in some more fundamental measure for degrees of freedom. Does this separate the symbolic base (for requisite variety) and the symbols (for information)? Consider how to measure this for Conceptual Spaces and Semantic Vector Space models (e.g., number of ANN parameters).

²⁹⁸ Research: There are presumably also subsymbolic-to-subsymbolic updates in a Knowledge System. It would be valuable to determine a way to isolate the four different theoretical update processes (i.e., subsymbolic-symbolic, symbolic-symbolic, symbolic-subsymbolic, and subsymbolic-subsymbolic) to assess their prevalence and uses in human learning.

²⁹⁹ More accurately, as will be discussed later, they do not ground the messages at the level at which they are predefined.

receiver to explicitly compare it against other possibilities.³⁰⁰ The third-person measure that quantifies models selected against possible models that could have been selected forces the measure to include measuring something theoretical and outside the bounds of the recipient. Doing so would be like determining how one will act by considering the ways one might act. It is a useful external measure but does not describe the fundamental properties associated with the process of evoking actions for the agent. External processes can indirectly constrain propensities for actions, but this differs from directly constituting action processes by them. If we want to measure the Semantic Processing *intrinsic* to agents, then we need a measure that can be constrained by evolution or subsequent effects but does not explicitly require them as reference points, since they are not directly applied by the recipient during the actual semantic processes.

Counterfactual measures also create the intractability common to many current modeling approaches, since they must explicitly model the system they are measuring, which is not foundationally necessary (e.g., Darwinian evolution does not need to model its own processes, even if it forms agents with modeling capabilities). Natural systems do not fundamentally explicitly compute/model how symbol sequences update some probability distribution over possible symbol sequences. There is no evidence that implicit semantic processes involve some explicit comparison against other possible interpretations. While a “possible worlds” measure can be used to optimize some model from an external modeler’s viewpoint (who receives the model as its own message), what we fundamentally need for AI is an intrinsic measure.

Symbols Cannot Emerge Semantics

A counter to the distinction of Semantics Processes from Shannon Information is that the latter can model a mapping between two levels (i.e., two different symbol sets). Shannon Information is positive when the number of states of two coupled systems is less than the product of those to systems alone (Rovelli 2015). That is, there is some correlation between the systems. One could therefore seemingly measure Shannon Information with two different systems as macrostates, each with different alphabets. The same is true for Semantic Information.

These alphabets can include letters to words, words to phrases, words to animals, pixels to animals, etc. However, deeper consideration for any models engineered to process information across such symbol sets will find that there is still an external modeler determining how to map across those levels. This coupling process converts what seem like two separate symbol sets into one set. Such cases do not count as two systems, but one larger system, as the symbol set can be

³⁰⁰ Of course, there are many cases where we explicitly compare our own predictions in the first-person as modelers. I do not preclude receiver’s from explicitly reasoning over possible worlds. A receiver can form concepts to explicitly model possible interpretations and choose the most likely one or seek additional information to eliminate possible interpretations. Such estimations are important processes that we as receivers undertake but are not the fundamental processes we are looking for in an absolute semantic measure. Instead, they are a derivative process that itself first requires forming a concept to model the data from which to reason about counterfactuals. What I am ruling out is a more fundamental counterfactual process. Explicit counterfactual reasoning makes up an important but tiny subset of human semantic processes, and it would be a mistake to project the assumption that it is a foundational way to model intelligence or assume a model is doing this implicitly all the time rather than being an extrinsic emergent concept applied by an external observer.

associated with symbols plus any *explicit* transformation rules.³⁰¹ Knowing how to relate the symbol sets may require grounding, but in these cases, the grounding resides in the external modeler. Thus, without subsymbolic processes, there is in theory a single symbol set to describe both the inputs and outputs, thus negating the role for Semantics. Such models can be completely described by Shannon-like Information but have no Grounding.

Moreover, such a Shannon-like measure is not intrinsic to the system itself. For a system to be an Information Processing system requires an observer. For ERM, the observer is the human that creates the ERM. Models that do perform semantic processes intrinsically can also be said to process information intrinsically based on criteria I propose in a later section.

This distinction has implications for any claims that symbols alone are sufficient for emerging semantics. I cannot imagine how this would be possible, even if symbolic operations were applied recursively. All ERM methods involve humans setting labels to represent symbols that are linked to other symbols. The references that these symbols point to are not intrinsic to the ERM; they instead exist in the human's mind external to the AI. It is only when the system is defined to include the modeler that emergence is possible, and with it, symbol grounding (i.e., semantics). This matters for AI: The Intelligence we know of is associated with emergent experiences related to symbol grounding. This grounding is essential for what results from "operating" with symbols. What is remarkable, then, is that the intelligence that has been built with ERM differs completely from the type many would envision for AI based on biological intelligence. ERM simply complete the calculations, without any understanding of what the symbols mean to external modelers. ERM require an intermediary, the human, to interpret the symbols. This is not true for natural systems that evolved intelligence, where, whatever it is, evolved through interactions with the world. As I have already stated, **ERM are models, not modelers. For a model to also be a modeler requires the ability to construct (i.e., form) models (i.e., micro and macrostates).**

Making this mapping probabilistic does not change this fact. **Consider how scientists make discoveries. They determine in their heads what tests will suggest relationships between variables; they choose controls/interventions by assuming how variables relate. Then they run statistical methods to test those relationships, such as correlation measures, clustering, classification, and/or significance tests. My claim is that ERM are more like how those statistical methods are traditionally used. The design of the probabilistic ERM, which is performed by the *human* modelers (i.e., determining which relationships to estimate values for between variables) is like the human scientist choosing the scientific experiments and hypotheses. Once the modeling decisions are completed by the human, the rest of the process is simply automated computing over the evidence, including given new evidence from the environment, where humans have already determined how such data will map to the ERM to be automatically applied as updates.**³⁰²

³⁰¹ Note that it is precisely because these would be limited to explicit rules for models that, e.g., map pixels to animal classes, would not work very well. There is no reason to believe an effective explicit model that symbolizes both pixels and animal classes would exist.

³⁰² In this sense, calling ERM "intelligent" seems like calling a song "musically talented".

Symbol-to-symbol manipulation is ungrounded and thus only works (when it does) because we, the modelers, used our own grounding to base our assumptions about what symbols manipulate other symbols and under what conditions (again, based on other symbols). This creates brittleness when the choices we make for how symbols affect each other artificially narrow the options that grounding could otherwise cover. Grounded systems can provide an endless set of possible models and associated inferences that cannot be emulated in a model that bypasses grounding. Thus, Grounding is not just critical but equivalent to Semantic Processes, which are intrinsic.

The above claims also should not be taken to imply that Semantic Information is useless. There is clearly a difference between a model for the probability of letters, a model for the probability words, and a model for the probability of events in the world that words represent. What we can say is that Semantic Information and Shannon Information descriptions generally occur at different levels. However, these processes all involve operations over symbols and are thus Shannon-like, in the sense that they do not address how the symbols are formed (i.e., how message elements come to reference something). **Any process that starts with an explicit micro and macrostate can be described as a Shannon-like model over already determined interpretations. Information models operate over the *results* of interpretations. Anything at a symbol level can be measured as Shannon-like Information, and the Semantic properties are the accompanying subsymbolic grounding for the symbols. The role of grounding as forming models involves a subsymbolic base from which symbols emerge given messages³⁰³ to evoke and be evoked by those models.**

This, of course, entails that the Knowledge System described above must be subsymbolic, with the ability to decode symbols and encode (emerge) them from the subsymbolic grounding of which it consists.

Finally, although I do not discuss Syntax in this chapter, there are theories that Semantics can emerge from Syntax.³⁰⁴ Instead, I believe **syntactical rules places structural constraints on messages, which help to reduce possible interpretations of message symbols.** Syntax, thus, constrains how semantics are grounded. This can be as simple as a period at the end of part of a message to indicate a new sentence has begun or a comma to separate two related but distinct assertions. Syntax, therefore, plays a role in reducing the degrees of freedom for forming micro and macrostates to interpret messages. It seems that syntax must itself be grounded for a receiver to understand how to condition interpretations of received messages. If so, then Syntax itself requires Semantic Processes³⁰⁵ as it involves symbolizing message elements, such as

³⁰³ This process was exemplified in the way distinctions on the face of dice can be symbolized. In the same way, a “.” can be a colon, two separate symbols for points, etc. The object that produces the message here is constant but can be symbolized in different ways.

³⁰⁴ For example, Jerry Fodor claimed that syntax can be used to imitate semantic process, where syntax plays the role of mediating between symbolic processes and their content (Fodor, 1985). If true, then it is not impossible that the entirety of AI is merely such imitation. However, this raises the question of how the content occurs at all. My claim, of course, is that Semantics is used to determine how to map the symbols to a model in the first place.

³⁰⁵ *Research: There is no doubt that syntactic changes can change semantics. A misplaced comma can lead a receiver to different interpretations (“You are, Dad!” vs. “You are Dad!”). However, the same concepts (e.g., “Yes” as an*

punctuation. It can also be implicit given messages, such as the ordering of word types (e.g., placement of verbs and nouns). At the very least, syntax does not seem more fundamental than Semantics, as there are often exceptions, and Semantic Redundancy (discussed later) within messages can often correct syntactic errors, such as missing commas.

Grounding as Model Forming: Prerequisites

With the above defined, I can begin to lay out properties to characterize Grounding or what I interchangeably call ‘Semantics’ or ‘Semantic Processes’.

Grounding should be defined and measured with respect to the message received and how it interacts with a receiver’s Knowledge System.

Since Grounding involves subsymbolic processes, knowledge should be implicit at its core. It is shaped by what has been learned over that receiver’s lifetime (including near-term) and what constraints have been adapted for over the receiver’s evolutionary history.

The Knowledge System provides the context with which models are formed. From the Knowledge System, Grounding is able to form different microstates and macrostates for different messages or even the same message given different contexts. Any system with such Grounding is not just a model, but a modeler.

Grounding should be associated with how messages are interpreted (i.e., get mapped to some set of symbols), which means that they evoke certain properties in the receiver’s Knowledge System. This is generally consistent with the intent of most existing Semantic Information Theories.³⁰⁶

Since grounding of messages depends on knowledge, Semantics depends on the receiver. **The external environment, including messages (i.e., data) received from it, does not inherently possess semantics. It is only in connection with effects to the observer’s Knowledge System that semantics are generated.** Trying to determine some fixed semantic properties of messages themselves leads to many issues, including how to ever observe them independent of the observer’s model. Recipients with different experiences, neural architectures, and even contexts, have different semantics evoked for the same messages. It is Effective Processes that deal with how useful the message was, which in turn shapes the semantics evoked in the agent for the

affirmation vs. “Yes” as referring to the progressive rock band) can have different meanings without any syntactic differences. One might argue that the syntactic properties must be extended to somehow account for how the brain structures these different concepts, depending on the given context. This seems to be addressed by considering the importance of reference frames discussed later in this chapter. I therefore will not explore it further, but it is worth considering how syntax and semantics interact across reference frames. It is also worth study to better understand the relation of Syntax to Conceptual and Semantic Vector Spaces, as well as how it developed and what is unique about how it can significantly constrain model selection.

³⁰⁶ In a later section, I describe how model forming can be measured proportional to how much the formation reduces possible interpretations about some state of affairs given messages. For Grounding, I want to swap out the notion of “reducing possible” interpretations with a process that does not rely on counterfactuals. As already stated, Grounding should be measured as from the perspective of the given observer. The external impacts of modeling “decisions” absolutely constrain what models a receiver tends to form, but it is not a necessity for them to measure/compute this for model formations given a message. Receivers can model the process of measuring this, but this is a (rather sophisticated) model of the more fundamental modeling process I am describing here.

message.³⁰⁷ Because we evolved over largely the same lineage, and because we experience relatively similar environment mechanisms, we are likely to experience similar grounding in many cases. If there is any functional value in a particular grounding, then it is likely similar for all humans able to perceive it since any grounding adaptations within species would have been constrained by their effectiveness.

Data only has potential for information with respect to what it actualizes for a given model. Data has little potential information for some models and high potential information for other models. The distinctions in the medium we call data may have some upper bound for potential information but no lower bound. These bounds are constraints on information, but they cannot be called Information, let alone Semantics, independent of the model impacted.³⁰⁸

Semantic-Level vs. Information-Level Interpretation

Associating Semantics Processes with interpretation does not readily distinguish it from Information Theories, since Semantic Information Theories are associated with reducing the number of possible interpretations. Likewise, Shannon Information can be said to interpret what a message means for a microstate through the changes it evokes in the macrostate (e.g., a sum of '4' for two six-sided dice outcomes = two possible microstates: 3+1 or 2+2). The difference between Interpretation for Semantic Processes and Information-based Interpretation links back to the difference between Percepts vs. Concepts. I will elaborate on this later. The point to remember here is that Grounding enables all (intrinsic) Information measures by giving rise to the models against which Information is measured. Semantic Processes interpret messages by determining what models to form. Information then involves interpreting between symbols *within* that model. For Semantic Information specifically, this is typically associated with some probability distribution over interpretations of a message. However, the possible interpretations are contained within an already formed model. A consequence is that an Information measure for reducing interpretations is counterfactual, while Interpretation in the Grounding sense must be (fundamentally) noncounterfactual in order to be an intrinsic and subsymbolic measure.

Unlike particular Information Theories, Interpretation in the Grounding sense is not scoped to a particular Reference Frame. I extend an example from (Bao, et al. 2011) to make this clearer. Say a receiver overhears someone say "the ___ machine in the kitchen is broken." Based on sound alone, the blank could have been 'coffee' or 'copy'. Interpretation between the two is intuitively semantic, as it requires knowledge to select between them. For "coffee" vs. "copy", one can in theory imagine some construction of candidate models that can be weighted, the computations of which can be explained as a Shannon-like Information process. For example, this could be a Markov Model that compares the probability of the symbols 'coffee' and 'copy', not only given the surrounding symbols in the message received but also probabilistic relationships with

³⁰⁷ Note that the receiver's knowledge can add (or reduce) complexity to messages that was not encoded by the transmitter from their own (theoretical) reference frame. Thus, the transmitter's requisite complexity (conceptually speaking) can differ from that of the receiver.

³⁰⁸ This goes back to the earlier footnote that there is inherent difficulty with defining data without assuming a model, but that we can still negatively define data as that which influences models but is itself model-less.

symbols for whether one is in a work environment, a symbol for the last known operational status of different appliances, a symbol for the time of day, a symbol for the transmitter's propensity to use different appliances, and so on. While some might equate a Markov Model to performing semantic processing, there is no fundamental difference in measuring uncertainty over a model with elemental symbols in the message vs. symbols that represent a large number of real-world objects and events. Uncertainty is captured in both cases using Shannon-like Information. What differs in these cases is what the symbols represent, which is critical to what Information model is formed. This is what Semantic Processes do. Even if I do not know whether the transmitter said 'coffee' or 'copy', knowing what these terms mean seems more semantic than having never heard them before. This involves forming a macrostate at a level of word meanings (at the dictionary level), which differs from the macrostate for understanding what words the transmitter sent in the message, as well as the macrostate for the state of appliances in the kitchen area.³⁰⁹ My claim is that these all involve Semantics. Some levels may be more common and semantically rich for humans, but they all require interpretation. Semantics should generalize to any reference frame that emerges symbols from subsymbols, not just those about concepts in our world.³¹⁰

The example above also demonstrates that truth conditions do not distinguish Semantics from Shannon Processes. As Shannon Information does not inform on the meaning of symbols, veridicality indeed appears to elude it. However, I believe that 'truthfulness' is one of many possible grounding frames (or attributes), and it would impact the Shannon-like model formed. Many message types are grounded without truth properties being grounded with them. More generally, it seems self-evident that grounding/semantics does not *directly* depend on some external right or wrong, or more or less effective, measure about the selected properties for the given symbols.³¹¹ For example, an agent that grounds the symbol 'Stop' with properties associated with moving forward is still grounding. This avoids arbitrary objectives from outside the agent being required to determine how much of something is going on intrinsically for it. Grounding therefore has nothing directly to do with "truth", except in cases where the grounding underlying models emerges explicit truth properties; and this is no different than any other symbol.³¹²

With all that said, while 'Interpretation' is an appropriate description for what Semantic Processes do, I replace it with the concept of evoking some degree of **Semantic Specificity** to

³⁰⁹ Note that there is nothing stopping the formation of an implicit model that actuates information across these levels. Again, the use of explicit concepts in the forming of micro and macrostates in this manuscript is primarily for explanatory purposes.

³¹⁰ We must also keep in mind that a system's level and quality of semantics is different than how semantic they seem to external observers.

³¹¹ I have claimed that Effectiveness does constrain Semantic Processes, and this would certainly include some capacity to assess the veridicality of statements. What I am claiming here is that the Semantics evoked by a message is not directly assessed by some counterfactual measure of its effectiveness or truth. As an analogy, hours of practice shapes my ability to play the piano, but a given piano performance is not directly measured in unit of practice hours.

³¹² This why I believe formalizing and measuring what is known as Semantic Information is less interesting than formalizing and measuring Grounding.

avoid conflating it with a counterfactual process and Information-level Interpretation. I elaborate on what I mean by Semantic Specificity later.

Information is Symbolic, Semantic Processes are Subsymbolic

There is a lot to discuss, so I want to stop for a moment and provide a concise explanation of the primary claim I will be making about Semantics and why I make this claim. Think back to the ‘coffee’ vs. ‘copy’ machine example and how understanding the words surrounding ‘coffee’ and ‘copy’ increases the probability for one interpretation over another. **The standard take is that the probability of the symbols (e.g., words) is a Shannon Probability, but what those words mean and how they therefore modify some understanding of the world is a Semantic Probability. But this alone has problems: It does not explain how meaning occurs. It only assumes it, and this is assuming the very problem for which Semantics is relevant. Understanding what “mean” means is at the very heart of semantics. Any Information measure ignores meaning, not just Shannon Information. In fact, Shannon Information can be measured against any macrostate consisting of some probability distribution of some index of symbols. This is not just a conjecture; it is common practice. In the coffee vs. copy example, the Shannon macrostate could be related to the probability of the word sequence, the probability that a coffee or copy machine was referenced in the message, or the probability that one of these is broken in the kitchen area. It does not matter if the probability is logical or statistical. This is not readily apparent because different choices of symbols present different semantic grounding, and some *are more common to us* than others. These problems help to understand what we need from Semantic Processes, which form micro and macrostates (i.e., models). Model formation determines the grounding of the symbols in a message. This is why I emphasize Semantic Processes / Grounding and not Semantic Information.**

The best way to understand the role of Grounding is to consider how impossible it would be to ground symbols one does not understand. It cannot be mapped to other macrostates about some worlds.³¹³ Concepts like Coffee and Copy are grounded in that the symbols ‘Coffee’ and ‘Copy’ have some underlying subsymbolic meaning for a receiver. They implicitly map to phenomena in some domain(s). Now consider the coffee vs. copy example but replace all the words in the message with symbols that are meaningless to us but still meaningful to other receivers (e.g., terms for coffee and copy from an unknown language). We find that this can be fully described using a Shannon-like Information model. It is only when we consider how the mapping between the systems was defined by the receiver to allow the updates to occur that the Shannon measure seems incomplete, and of course this where I claim we find the Semantic Processes. More precisely, **any example with messages composed from the same symbol set as the macrostate updated by them can be comprehensively measured using Shannon-like Information.**³¹⁴ There are many possible models that can be used as a symbol-level state space to associate with a message, including a model about the interpretation of the message itself.

³¹³ I will reiterate this point in later sections, as it helps keep what grounding does in focus.

³¹⁴ This is even true for the truthfulness of the message. That is, it can in theory be demonstrated that the truth value for assertions can be added to a symbol set and then measured using Shannon Information. I tentatively claim that

Thus, Semantic Processes determine what a symbol will symbolize, and hence the meaning of symbols.³¹⁵ However, once symbolized, there is some Shannon-like model that can associate those symbols as part of the state space. This is exactly what we do as modelers for ERMs. We replace the role for Semantic Processes in the AI itself by using our own to form the models. All Semantic Information Theories I have come across seem to do this too. Without modeling Grounding processes to emerge the symbols, AI will not have the appropriate representational structure to learn the semantics itself, greatly limiting the significant role that semantics has in developing intelligent systems. It is difficult to see this issue in Semantic Information theories because we ground the symbols ourselves as external observers to the example (that is, we are not the receiver in the example, but we project our grounding on the example we are observing). These Semantic Information Theories are starting with the symbols that would emerge from grounding, rather than the grounding itself.

The role for Semantics is grounding (interpreting) message symbols using a Knowledge System. Since this actualizes the Shannon updates of that Knowledge System, it also changes the grounding itself. In other words, the relation between subsymbolic and symbolic processes goes both ways. Symbols manipulate the grounding. **This demonstrates another issue with the Physical Symbol System Hypothesis which assumes only that symbols manipulate each other directly.** I might argue that the brain is essentially a grounding system. It is impossible for it not to ground stimuli in some way if experienced. For example, a message in a foreign language will not be grounded in concepts of the world as they would in our native language. However, the message may be grounded in phonemes heard and/or the perception of not being able to understand it. This is not to say that all grounding across reference frames is created equal, but it suggests that any experience is a grounding of something, and we do experience receipt of messages in unknown languages. Thus, it would be more accurate to say that we understand such messages but not in the sense that one would expect or desire to understand. The understanding occurs within a different reference frame. **Choosing how messages “should” be understood should be measured as an Effective Process,** which is discussed later in this chapter.

Criteria for Subsymbolic and Intrinsic Processes

Given the above claims, it is imperative to define what ‘subsymbolic’ means, as well as what can be claimed to be *intrinsic* processing of messages. I claim that these two are fundamentally

this proves that truthfulness is not fundamental to Semantic Information. However, nothing is fundamental to Semantic Information. We can choose to define it however we want, as long as we recognize that it is the Grounding process that is really the core to Semantics.

³¹⁵ This may seem odd, since the model chosen seems to itself determine what the messages mean. However, in the case of explicit micro and macrostates, I believe this is best thought of as what is evoked in us as extrinsic observers. This is natural, since symbols can be transmitted, grounded, and evoked again within and across modelers. However, as I mentioned elsewhere, the distinction between the grounding and (implicit) micro and macrostates may be more appropriately considered a conceptual device to better understand semantics and intrinsic information processing. It is a means to isolate what is missing in our theoretical understanding of Semantics and Information given limitations of using symbols to describe it all. In fact, one could argue that the grounding of the symbols is equivalent to the micro and macrostate formation, rather than one causing the other. However, it often seems clearer to describe them as separate but related to understand the points made in this manuscript.

related. In this section, I propose criteria to determine what phenomena should be considered subsymbolic and intrinsic. Some of the assertions should be viewed as initial theorizing that will be further developed beyond this manuscript. Still, I believe the section describes critical properties towards such theories.

My claim has been that grounding involves forming models (i.e., micro and macrostates) given messages, where the symbol set formed differs from the processes that form it. To be Intrinsic simply requires that the system of interest emerges the symbols itself. From the perspective of some chosen level at which we place the receiver, the underlying processes are outside of the symbol set for the micro and macrostate this level forms. **To be outside the symbol set is also necessarily to be subsymbolic intrinsically for that system.** This requires that the system dynamics cannot be explained by the symbol set it forms (i.e., the Shannon-like macrostate and corresponding microstate). There must be some level, such that, from the perspective of that level, the underlying processes that form it are subsymbolic and ground the symbols encoded and decoded. This breaks the chain of empty conceptualized symbols. To explain this criterion, I will use Conceptual / Semantic Vector Spaces and contrast them with ERMs. **An underlying neural network is subsymbolic to the Conceptual Space it forms. Through the neural network dynamics, the Conceptual Space grounds symbols for received messages, where those symbols make up the index of the formed micro and macrostates. This is required for any information processing done with those models to be considered intrinsic to the system. Otherwise, we fall into the trap of conflating extrinsic measures with measures of intrinsic processing.**

However, it turns out to be difficult, given our limitation to an extrinsic emergent perspective, to provide both satisfying and observable criteria. There is no guarantee that we can even qualify something external as involving intrinsic processing. Defining something as ‘intrinsic’ is tricky because we as external observers can always eliminate any model to its end-to-end mechanisms (e.g., bits or atoms), making any assessment for emergent symbols extrinsic emergent.³¹⁶ That is, one could argue that any digital model can be described as taking in bit level data, processing bit level data, and outputting bit level data. If any digital model can be described end-to-end at the bit level, then is there really any grounding difference between ERMs and ANNs? Neither would seem to meet the grounding criteria.³¹⁷

³¹⁶ Note: As third-person observers, our extrinsic assessments for information processing of external systems occur as properties within our own intrinsic processes. These extrinsic assessments are derivative of our own intrinsic processing since only some of our intrinsic processes would be associated with our modeling/measuring of extrinsic processing of external models/systems. Thus, it also fails as a measure for our own (Information, Semantic, etc.) processes. Distinguishing extrinsic from intrinsic assessments of information processing is therefore not only important for AI but also for understanding something core to our own cognitive processes, including how we understand (or construct) the reality we perceive.

³¹⁷ Another way to put the challenge is that a measure for something subsymbolic would seem to necessarily conceptualize its variables as symbols, which in turn makes any emergent symbols eliminable to its mechanisms. This links it back to the Percept vs. Concept distinction. Anything symbolized is conceptualized.

And yet Semantics is not epiphenomenal.³¹⁸ It has unique functional roles (e.g., semantic compression, analogy making, creativity), which can be demonstrated in various non-ERM models, such as Conceptual Spaces and Semantic Vector Spaces. It is core to intelligence as we know it; the symbolization process is produced from it, which only represents the tip of the iceberg for how inference is performed.³¹⁹

To summarize, Semantic Processes must be measured intrinsically and subsymbolically. The challenge with describing a measure of what is intrinsic to the system of interest is that we conceptualize what is fundamentally a perceptual process. The related challenge with the process being subsymbolic is that one can conceptualize an end-to-end process using a single symbol set for such systems, such as bits for a digital model, without any room to incorporate emergent processes outside those symbols, except for those that are evoked in us as observers and projected back onto the system. Below are the criteria I put forward to address these challenges.

What counts as subsymbolic must address how a system could not be fully explained end-to-end at the level of bits. **My tentative criteria for being subsymbolic are that we, as third-person observers, must be able envision a reference frame for a system with a different symbol set than the underlying processes that give rise to it, and that these underlying processes determine this different symbol set. A reference frame should be thought of as a first-person perspective at some given emergent level.**³²⁰ **There must a particular type of dependence**

³¹⁸ Indeed, an argument that might be used against Semantics being subsymbolic is that it affects probability distributions of words. That is, to be semantically coherent, words cannot be randomly distributed or even just constrained by grammatical rules. However, I believe this is an effect of semantics rather than the distributions explaining semantics. This is also why a high-order Markov chain cannot ground words as symbols like a neural embedding method does. While the latter learns from probabilities of words based on occurrences, it does so through a subsymbolic medium. Thus, statistical patterns can be leveraged to learn semantics but only if the receiver learning through statistical patterns meets the subsymbolic criteria proposed in this section. Moreover, those statistical processes are a conceptualization of the underlying dynamics that ground symbols at a different reference frame.

³¹⁹ Much more speculatively, Semantic Processes represent what of Reality is only indirectly captured in the Sciences, *even non-reductive fields*. All Sciences deal in explicit symbols for emergent phenomena. The symbols ride on the subsymbolic grounding that resides in us intelligent agents but cannot be directly communicated because doing so converts them to lossy Shannon Information descriptions of the world. All current Sciences conceptualize the phenomena they describe, which become eliminable when they are the object of study. I believe what motivated the emphasis on nonreductive phenomena was the seeming missing grounding in reductive phenomena (that which cannot be eliminated). If the phenomena are conceptualized, then the problem that drove the focus to reductive phenomena is no more solved for symbols representing higher level phenomena than it is for basic symbols. The “nonreductive” ones, which represent concepts for things or process of some world, just happen to evoke more semantically rich phenomena in humans. Intrinsic emergence itself is this subsymbolic content, which is an explanation for why it is difficult to understand how emergence arises within a scientific theory, at least as they have been developed. If this is true, then AI is holding up a mirror to Reality. It shows that Science (or any conceptualization process) is limited to a symbolized and therefore conceptualized and eliminable (in principle) form which is derived from a comprehensive Perceptual/Semantic grounding. Fortunately, we as humans are grounded and thus can decode and encode this derivative symbolic form. However, there is no guarantee that all Perceptual processes can be derived to and through symbols. Whether there are means to avoid the limits of conceptualization and remain scientific is worth future consideration.

³²⁰ More accurately, a Reference Frame can be thought of as the intrinsic version of an emergence level.

whereby the processes determine the emergent symbols rather than choosing from some predefined symbols or symbol sets. Otherwise, the underlying process is operating over the same symbol set at the emergent level we envision.³²¹ To be subsymbolic, the level for the underlying processes can be symbolic to an external observer, but they cannot be symbols to the model at the level in question.³²² These claims seem warranted when considered from the perspective of a receiver at this given level, since these underlying processes need not be modeled (i.e., symbolized) by this receiver for the micro and macrostates to form given messages. If they did symbolize something, that would require something to interpret them, creating an infinite regress of levels to map across.

Of course, us envisioning a Conceptual Space as a reference frame is still extrinsic emergent because it is us conceptualizing the reference frame for something else (i.e., the reference frame for the system envisioned is a property of our own reference frame). However, we can ask questions about such a *theoretical* Conceptual Space as a reference frame. **That is, we place no initial requirement on it existing digitally or physically. In doing so, we find that an envisioned Conceptual Space is not end-to-end explained at the same symbolic level. From the reference frame of the system itself at the level of the Conceptual Space, there are no symbolized bits, neurons, etc.**³²³ This is not because such processes do not emerge the Conceptual Space but because symbolizing them is not required to form macro and microstates. Likewise, when humans ground symbols, they do not need to symbolize the neural operations that give rise to and ground those symbols. **The symbol set that emerges, emerges at a different reference frame (the Conceptual Space) than the symbol set we as external observers would conceptualize for the ANN (e.g., nodes, edges, weights) or at the level of bits for underlying digital code (i.e., 0s and 1s).**^{324,325}

Thus, micro and macrostates formed from Conceptual Spaces (e.g., possible Twenty Questions answers) are grounded because the neural network dynamics that emerge them are

³²¹ We will see that another way of saying this is that, for a given message, Semantic Processes and Shannon-like Information are measured at different levels. For Semantic Processes to be distinguished from Shannon Information, there must be constraints outside the formed model that impact what model was formed. Based on the criteria described in this section, the distinction between Semantic Processes and Information is a fundamental one for any given Reference Frame (i.e., first-person perspective).

³²² Note that, what is subsymbolic to one receiver need not be subsymbolic to another. What matters is that the AI must determine what messages mean, which requires forming the symbol sets through which to symbolize what features are represented given the data. We will see that this ability has functional consequences, particularly for achieving AGI.

³²³ There of course could be, for certain models, percepts that include concepts of bits, neurons, etc. along with concepts of chairs, dragons, statistics, etc. But these are emergent symbols; not the underlying layer being considered here. These conceptualizations are not necessary for all micro and macrostates that could be formed. In fact, micro and macrostates that include grounded symbols for neuronal properties are just as divorced from the actual processes that give rise to them as any other symbols would be. This relates to both the Description Gap and the Explanatory Gap.

³²⁴ This is why Reference Frames are important for understanding Semantics.

³²⁵ **Research: What about methods like Factor Analysis? When are factors (i.e., latent variables) grounded? If they can be grounded, then are explicit variables, such as observables in Hidden Markov Models grounded through the dependence on hidden variables? If so, then in what ways?**

subsymbolic at the level of those states. If the reference frame was at the level of the neural network, then the neural network properties would make up the macrostates (e.g., prediction of what neurons will activate given upstream signals). At this level, the neural updates are symbolic. Messages could consist of neural activations, weights, etc., or weight vectors or matrices, and can therefore be modeled as Shannon-like Information. However, envisioning the neural network processes mapped out as purely symbol transformations is not taking the theoretical model's perspective. That would be a theoretical model with a different reference frame. Likewise, at the machine code level, the inputs and outputs that are bit strings to us are not bit strings to the theoretical Conceptual Space model. Thus, the difference between Grounding and Shannon-like Information depends on the reference frame and whether there is some level that is subsymbolic given that reference frame. The reference frame level would have symbols grounded in subsymbolic processes.^{326,327}

Note that it does not matter that the concept that there is some symbolic level that emerges from the lower level is extrinsic emergent. We are not claiming this level actually exists. It is a theoretical system and therefore a theoretical measure: It shows that, if this reference frame did exist for the theoretical system, this is what it would be like for that system to intrinsically measure processes and have grounded symbols. This can in theory be done for any arbitrary levels we choose to conceive of, but only some will end up being good empirical fits for how a given AI model in question behaves. Therefore, we can set the theoretical model, determine if it meets our subsymbolic/intrinsic/grounding/semantic criteria, and then assess its empirical fit with the behaviors observed for specific AI models.

What would empirical evidence look like extrinsically? I think one possibility is observing a Description Gap when crossing reference frames. For example, what are symbols to us (ASCII symbols or pixel values) are lost once fed into the neural network. They instead contributed to the distributed representation of correlations among these symbols. Here, **our emergent view does not cohere with the envisioned theoretical system's reference frame. This would suggest**

³²⁶ This is a bit complicated, so I repeat the process being claimed here for a theoretical Semantic Vector Space: The ANN processes evoked by an input (e.g., one-hot vector that represents a word to us), while symbolic to us when we consider it at the level of neural dynamics, including the output layer activations, form an implicit micro and macrostate that is updated given the input. This emergent micro and macrostate, can be envisioned as symbolic for the ANN itself (i.e., intrinsic information) and, depending on the training (including other modalities than just text) and potential of the subsymbolic architecture, can in theory be similar to what we observe as words formed from a Semantic Vector Space. Thus, the underlying ANN from the perspective of the envisioned emergent semantic level, determines the indices used to update the word-context prediction model. At the envisioned semantic level, the ANN would not perceive the ANN level symbolically any more than we experience our own neural process as symbols when they ground messages.

³²⁷ I reiterate that I am not making a claim here that any current ANNs ground symbols the way we do (the way they must be grounded for AGI). I am only claiming they meet the criteria for grounding (to the extent that they empirically fit the theoretical model). Tests to measure the grounding properties for models must be done to determine in what ways they must improve, which would be expected to be significant given the need to address other claims made in this manuscript and elsewhere.

an alignment of sorts for the system we are extrinsically assessing with the theoretical model envisioned.³²⁸

To be clear, we cannot claim with any certainty that the theoretical system intrinsically *experiences* anything given these decoding and encoding processes. The results that we would associate with the emergent symbols are extrinsic emergent to us, just as would be the case if we were to reference the underlying neuronal dynamics of other humans or any animal. We have no access to the intrinsic experience of the symbols generated.³²⁹ However, I claim we can still say that, were such symbols in a position to be intrinsic, they would be grounded/semantic. Moreover, **as a semantic process, the updating of macrostates formed is an *intrinsic* Shannon Information process because performing actual semantics is the criteria for symbols to be symbols to the system itself.** A more controversial³³⁰ claim that relies on the functional definition I used for Grounding is that such a system, can be said to actually ground symbols, independent of any intrinsic experience, with certainty proportional to the empirical evidence.³³¹ I elaborate on the implications for this claim later in this section.

This approach does not seem to have all the issues I raised for conflating extrinsic with intrinsic emergence. Although an extrinsic emergent assessment, it is constrained by the need for there to be some theoretical model that, were it to exist, would, from its own perspective, meet the subsymbolic criteria from which the emergent symbols would arise. That is, it is not

³²⁸ *Research: Consider how these criteria can be used to assess various modeling properties. How do pre-modeled architectural constraints affect the semantics intrinsic to the AI being modeled? For example, Convolutional Neural Networks (CNNs) (Krizhevsky, Sutskever and Hinton 2017) have predefined constraints in how they convolve and pool features to build up increasingly abstract features as hidden layers get deeper. Thus, humans have determined certain constraints to guide the algorithm to learn in certain ways, such as being relatively translation invariant. Crudely speaking, according to the criteria in this section, is the CNN unable to ground certain features driven by these specific predefined choices? Would it be specific to those constraints? That is, would the AI not learn any semantics underlying pooling or convolution? Does this mean that the CNN could not ground concepts like feature abstraction itself and translation invariance? Note that these are not the concepts that we would assume they ground (i.e., what we associate with the labels with which we train them). Determine an empirical test for this hypothesis. If grounding is functionally adaptive, then we should be able to detect its effects on the features/concepts (implicitly) included within the macrostate, whereas we would detect its absence for features/concept associated with predetermined properties; those that are directly engineered into them such as convolving, forgetting in LSTMs (Hochreiter and Schmidhuber 1997), etc. The many varieties of Graph/Relational NNs (Battaglia, et al. 2018) offer a potential opportunity to test this, as they incorporate relational biases, which act as structural constraints for how activity propagates through them. Relational Reinforcement Learning (Zambaldi, et al. 2018, Džeroski, De Raedt and Driessens 2001) provides an example, where the engineered inductive bias could be something like learning nonlocal relationships (e.g., object A is under object B), but the actual relations are learned, such as the idea of “key-ness relation to locks”. Here, nonlocality itself may not be grounded, but the key-ness relation to locks could be. The functional value is demonstrated, as the newly grounded concept generalizes to other environments without having to relearn this key-ness relation.*

³²⁹ While we are in the same position with AI, we have more potential access to and control over their cognitive processes to be able to assess them extrinsically. This increases the potential for the empirical approach to determining the appropriateness of claiming what AI models ground symbols.

³³⁰ It is only controversial in the sense that some might insist that Grounding be defined to always require intrinsic experience. I described in the Definitions section why I believe this is a less useful definition. The claim above this one holds regardless and is sufficient for AGI purposes.

³³¹ What counts for empirical evidence and how it should be measured still needs to be formulated.

simply what is evoked in us when we observe dynamics of a system. Another way to think of the difference is that **the subsymbolic criteria provide an extrinsic emergent measure of an intrinsic emergent process for a system. This differs from an extrinsic emergent measure about an extrinsic emergent (i.e., conceptual version of a) system.** An ERM, like a book, can evoke grounding in us but does not meet the subsymbolic criteria placed as a requirement of any theoretical model of a given system for which we would say is semantic.^{332,333} In fact, **given this method of starting with the theoretical model, we can then check if the model is “extrinsic emergent complete”.** There is nothing left for ERM symbols to do since the dynamics of the given extrinsic emergent level we start with explain everything (i.e., is fully interpretable, since there is no decoherence / Description Gap). ANNs on the other hand, are not extrinsic emergent complete at the same emergent level as ERMs (e.g., the level of world objects and events); the underlying ANN dynamics are doing something that is not fully explained at this emergent level. There is more going on than the symbolic level can explain at that extrinsic

³³² Research: (Smolensky 1988) distinguishes neural networks from symbolic AI by the former involving numerical rather than logical processes. Consider how the bits underlying an ERM vs. an ANN actually differ. Do similar symbolic concepts represented by these models share any constraints that force some similarity in bit sequences for ANN numerical processes? What about inferential logical processes?

³³³ One might claim that for ERMs (or even books), there could be some theoretical level hidden from us that could be described as a different symbol set, analogous to the role of a Semantic Vector Space for ANNs. I am unsure where such a level would be, but if it were possible, it would be arbitrary; no more constrained than any random mapping. ANNs, on the other hand, generate specific maps to features. They cannot be scrambled and still emerge the same features (Chalmers 1992). The underlying processes for ANNs have some relation to the emergent symbols formed in the Semantic Vector Space. Technically, ERMs could have some grounding; after all, ANNs are essentially matrices and operations over them, and ERMs can also be represented as some structure, independent of the labels we attach to them. For example, they can consist of a matrix of probability values or Boolean values representing whether each variable pair connects. However, whatever emergent phenomena that would be grounded in the dynamics of this underlying structure would have no association with a grounding related to the explicit concepts the ERMs are meant to represent. An underlying dependency between two nodes would not be expected to ground an ERM subgraph meant to represent the dependency between rain and wet grass. There is no reason to believe this would change by adding many more symbolic variables and relationships as “context”. Any grounding these structures would form would on average have no more semantic relation to the explicit concepts used for labels than randomly generated ERMs. ANNs, on the other hand, are trained directly on the underlying structures and structural processes to emerge the symbols we expect them to, even though this is assessed by us extrinsically. Of course, this is by design for ERMs. The point I am making is that this design possesses negligible potential for intrinsic semantic and information processing, and certainly no relation to the emergent concepts with which we label them.

From (Chalmers 1992): “...whereas such bits might be entering into a computation, they are not individually causally efficacious. Whenever they enter a computation, they enter as a chunk—for instance, the eight bits that implement the atom ELEPHANT are always manipulated together. For all intents and purposes, these eight bits form an atomic unit. They stand in a strictly implementational relationship to the computation at LISP-atom level, and so do not have any significance of their own. The eight bits merely form an arbitrary label for the LISP atom; any collection of bits would do as well. This contrasts clearly with the connectionist case. In a connectionist network, a representation is made up of a collection of activated nodes, but each of these nodes plays a separate and significant causal role. We could not replace this particular configuration of active nodes by another configuration without radically altering the causal role of the representation. The distributed representation, unlike the LISP atom, cannot be treated as an atomic object.”

As already mentioned, this has functional implications for grounding. If the distinction I am making here is correct, then we should see evidence that it makes a difference in what systems with or without it (or more or less of it) can achieve.

emergent level. These processes are therefore outside the symbol set and face the Description Gap.³³⁴

Since ERM's are predefined with respect to how their explicit concepts map to "symbols" received, the set of symbols that would make up the emergent set at an ERM's level do not emerge from a different symbol set.^{335,336} ANNs have a middle layer that implicitly determines how bits map to emergent symbols.³³⁷ Since they reliably evoke symbols expected from the theoretical reference frame as subsymbolic processes, they can be said to at least act as if they perform intrinsic processing. For ANNs in general, humans determine how raw data features map to the input node and edge level, but humans do not determine the next level of mapping, which is from nodes and edges to abstract features forming symbol sets.³³⁸

We can likewise consider how criteria for Semantic Processes are achieved by biological agents. Neurons are not grounded in the electrochemical signals they receive from each other or as evoked by the world. Updating electrochemical responses given electrochemical inputs does not meet the subsymbolic criterion at the level of neurons. However, if something like a Conceptual Space model empirically explained the behaviors of the system, then we could argue that, at the level of that theoretical Conceptual Space model, there is a strong case for grounding in the actual system of interest; namely, due to the neuronal dynamics, which are outside the symbol set that emerges given these underlying dynamics. For example, one can ask if the neuronal dynamics of a human brain can be represented in terms of the Twenty Questions symbol set (e.g., dogs, cats, etc.). It cannot, as the neuronal dynamics would consist of completely

³³⁴ The 2nd wave of popularity for ANNs in AI was largely driven by grounding-like concerns with ERM's (McClelland and Rumelhart 1986), at least in the functional sense (rather than as grounding defined with intrinsic experience). This manuscript, in also focusing only on the functional sense, attempts to help establish the criteria.

³³⁵ Note that no end-to-end system could meet these criteria at the level of its own reference frame. Recall that Information makes no sense to a system unless the system includes an observer. Thus, if a reference frame for some model is described at the level of bits, then it does not "process" the bits at all. For Information to be something processed intrinsically to a model, even for Shannon Information, requires this underlying Semantic Process. This cannot be intrinsic for end-to-end 0s and 1s or end-to-end anything. Thus, we must claim for a given AI model that either there is some other level where subsymbolic processes can be defined for it or the envisioned model does not meet the intrinsic criteria. The latter case is based on the fact that the idea of taking the AI's perspective for an end-to-end (i.e., single level) system does not work even in theory.

³³⁶ This is a good place to reiterate that the issue with ERM's is not that graphs in general cannot emerge semantics, but that the semantics cannot be measured at the same level as the symbols representing the graph, which is how ERM's are used. In ERM's, the symbols meant to represent concepts are not grounded at another level. There is considerable opportunity to consider how, for example, different graph topologies affect semantics.

³³⁷ *Research: Analog robots can have predetermined voltage thresholds, etc. that determine when to turn left, right, etc. Are these grounded according to these subsymbolic criteria? Unlike ERM's, these mappings are not all arbitrary. A different threshold would change their behavior. The difference is whether there are concepts modeled that depend on those input values. That is, it depends on whether those dynamics can be said to underlie phenomena that would be conceptually symbolized at another level that differs from but are caused by them. It is worth studying a variety of robotics experiments to assess the concreteness and utility of the claims in this chapter.*

³³⁸ Note that, if the above claims suggest that Semantic Vector Spaces capture grounding, then it seems that any separate models using its outputs for inputs may suffer from the same problem as ERM's, except instead of being parasitic on human grounding, they are parasitic on the Semantic Vector Space model. I believe the criteria would be that the Semantic Vector Space is involved in any symbol manipulation. If symbols manipulate other symbols without using the Semantic Vector Space as the medium, then that process involves no (or limited) semantics.

different symbols (e.g., neurons with synapses of varied strengths, resting potentials), which do not map one-to-one to the Twenty Questions symbol set.^{339,340}

Thus, **the way to meet the criteria for Intrinsic and Semantic Processes is to start from the reference frame of some assumed model that deals in a different symbol set from the underlying dynamics that evoke the model given inputs.**³⁴¹ If that model were to exist, then it would be subsymbolic. Whether we know the model exists “out there” (or “in there”) is a separate question. Namely, it is a question about the appropriateness (or usefulness) of it as a model for some system of interest. As for all models, this is determined by their ability to explain phenomena observed and to meet predictions, including seeming use of the symbols we expect and extrinsically assess given inputs. The difference in these cases is that the model being assessed is a modeler. Conceptual Spaces not only meet the criteria but are promising models for cognition in natural agents. Actual Semantic Vector Spaces seem to meet the criteria, which is consistent with their use for recent breakthroughs in AI.^{342,343} ANN-based Semantic Vector Spaces are not the only systems that could meet these criteria. The agents in an agent-based model or cell-level rules in cellular automata could be viewed as subsymbolic to some emergent grounded reference frame. None of this guarantees that any given system would have semantics similar to ours, but there is no obvious reason to deny that such systems would not ground something to the extent that they meet the criteria proposed.³⁴⁴

What this approach does is shift the problem of creating an intrinsic measure stripped of extrinsic emergent projections to a less foundational problem. Starting with some external model

³³⁹ For brains and ANNs, this is because representation of those symbols is distributed across the network.

³⁴⁰ Note that the operations performed at the lower level can be anything as long as the dynamics they create emerge the higher-level symbols and are outside that higher level’s symbol set. They can be Bayesian, physical equations, Shannon or some other Information measure (e.g., Integrated Information (Tononi 2015)), and maybe even logic. It does not matter if these are symbols to us as external observers, because those processes are not what would get grounded for the system of interest. Moreover, any such subsymbolic operations are in a conceptualized form in how they are described by us external modelers. This is not to say that any of the operations listed here are biologically plausible. They do not need to be to emerge *some* grounded symbols but almost certainly need to have more biological realism than current AI to achieve AGI.

³⁴¹ For the record, the inputs in themselves are just data, and thus model-less. Thus, a percept of them depends on the reference frame, which could ground them as bits, letters, words, etc. They are the messages that would be grounded.

³⁴² *Research: Even if semantic vector spaces are good representative models for how biological intelligence works, the way they are trained (learning by predicting co-occurrences of e.g., words in data) seems like more of a hack or indirect means to get at the same thing we learn in a different way. Determine a more natural but equally scalable way to learn Conceptual / Semantic Vector Spaces that could still potentially meet the grounding criteria.*

³⁴³ I will propose a nonexclusive approach for meeting subsymbolic criteria later that is specific to Conceptual and Semantic Vector Space models.

³⁴⁴ *Research: I realize these criteria leave many unanswered questions, particularly with respect to the second step of assessing if/how the results observed ground symbols and in what way. There is a need to formalize the criteria and measures for such assessments of systems of interest (e.g., ANNs, Cellular Automata, Agent-Based Models). For example, if the cellular automata for Conway’s Game of Life do meet the grounding criteria, then what would we be looking for to determine if this grounding is anything like the grounding we project onto the game as human external observers (e.g., gliders)? How can we compare how we ground the words associated with vectors activated in an ANN-based Semantic Vector Space vs. how the Semantic Vector Space grounds them?*

of a system of interest and trying to determine intrinsic measures always fails because there is always a causally sufficient end-to-end model at some level. Instead, we start with some theoretical model, take as our reference frame this emergent model itself (that which gets stripped away if we start with a physically realized model), and determine what is required to meet the intrinsic criteria (i.e., being subsymbolic with respect to the emergent macrostates it evokes). Then, the problem is pushed to the challenge of proving these intrinsic but emergent models exist for certain physical or digital systems of interest (e.g., biological brains, ANNs, ERMs). I claim this can never be proven because this theoretical model is extrinsic emergent for any of those systems; we would be projecting it onto those physical systems. However, just as for any models, we can then at least ask if the candidate model is empirically appropriate for the system of interest. If we determine it is appropriate, then we have not only determined it is appropriate as a model for that system, but a model of the modeler variety; one that does meet the criteria for having an observer that can process information intrinsically. Appropriateness can be based on predictive performance, explanatory power, etc. just as we assess other models in Science.

Implicit in this claim is that Semantic Processes are required for system to be said to process information intrinsically. This is important because others' theories that I am aware of for measuring intrinsic processes do not point to Semantics as playing this key role as a pre-information phenomenon. If Information requires an observer, and if Semantics is a process that

is necessary for an observer to be assumed for a system, then it makes sense to claim that Semantics is not a type of Information.^{345,346,347,348,349}

³⁴⁵ I contrast Semantic Processes with some Semantic Information Theories later in this chapter. This does not mean there is no utility in these Semantic Information Theories. Just because they do not capture how Semantics arise does not mean that they do not measure something useful. In fact, because Semantic Processes are more fundamental than any of these theories, it opens the door for any of the different theories to be nonexclusive measures that can be applied in useful ways.

³⁴⁶ Aside: The claim that all information is physical (Landauer 1991) is based on the fact that every concept, no matter how abstract, must arise from distinctions, distinctions require something distinct in stimuli (i.e., what evokes messages received), and stimuli must come from the world. I do not disagree (although see (Lopez-Suarez, Neri and Gammaito 2016)). This also justifies only describing Shannon Information and other symbolic measures as truly being Information. However, it is worth noting that this claim is made at the conceptual level. I already claimed that concepts are eliminable, which aligns with the assumption that Information, which is itself at the conceptual level, is ultimately caused by the lowest level mechanisms. The question, then, is whether Semantic Processes (and perhaps Effective Processes) also must be physical. For this, I would tentatively argue the question may simply be inappropriate for Semantics because the phenomena in question are fundamentally the Non-Conceptual Percepts and just “are”. Once we consider them as symbols, we have taken a different reference frame for them that does have the property that they must be caused by physical information processes.

³⁴⁷ *Research: More speculatively, could human dynamics be subsymbolic for some other level (e.g., society)? What are the implications for collective intelligence given these claims for emergence/grounding? Can semantics be achieved at the superorganism level (e.g., ant colonies)? If human dynamics are insufficient, then why? Integrated Information Theory (IIT) (Tononi 2015) proposes a reason based on a measure described later in this chapter. Consider building a single semantic vector space for a group of agents, such that their individual actions and states create the embedding by learning to predict what they do. This grounds vectors representing each agent's action and states, and perhaps rewards, as elements within each collective vector. Does this provide any advantages for certain domains over standard Multi-Agent Reinforcement Learning? It would choose or influence each agent's actions (when available). When unavailable, each single agent could have its own vector space to determine actions to take. These individual vector spaces could learn to predict what actions the collective vector space will choose (or even better, predict their positions in the collective vector space), to incrementally make the individual vector spaces more like the global one. This can be done in a single Graph Neural Network to coordinate multiple agents, each providing their observations and action confidences as inputs to the collective vector space. Thus, the model is centralized when the collective vector space is available and decentralized when not.*

³⁴⁸ *Research: Can information be the fundamental primitive from which the Universe is fundamentally built and still meet the requirement of having an observer? I do not see how, since then it is all symbols and no content. However, if we instead allow reference frames to be more fundamental than Information, then those primitive symbols could meet the subsymbolic criteria. If so, what are the implications for the notion that there must be a physical source to produce the distinctions for information to be available for a receiver? To say information is physical either means there must be a source that is as fundamental as the information itself (or more fundamental), or that information emerges the physical source, which would mean that physical stuff cannot be more fundamental than the information itself and the observer.*

³⁴⁹ One might wonder whether Semantic Processes choosing which models form could be described as a higher-level Shannon-like processes (i.e., a metamodel that chooses the models). Perhaps every Semantic Process can be described as a Shannon-like Process by focusing on a different level. Of course, if this is possible, then, at yet another level, selecting between metamodels given the message is another Semantic process, and so on (I doubt such iterations would be deep in real systems. Higher meta-levels might have increasingly reduced numbers of possible macrostate sets and the physical limits could stop the infinite regress). However, note that, since the Shannon-like level would have to convert the subsymbolic processes to symbols, the model would be of a different macrostate. Therefore, Semantics is still always subsymbolic, but what are symbols depends on the level at which one is observing the models. This of course simply aligns with my original claim about their fundamental differences. Thus, Semantic Processes are still necessary, regardless of whether one conceptualizes some nesting with Shannon-like processes.

One counter to criteria that systems cannot have all symbols predefined for it is that one could predetermine the weights of an ANN for a Semantic Vector Space. Specifically, one could train an ANN, and then use those weights to implement a new ANN with the same architecture. It would be absurd to claim that copying a trained ANN would remove the original grounding. That is not what is being claimed. The symbols at neural dynamics level are not the symbols at the level of the Semantic Vector Space forming models. The ANN weight matrix is the modeler at the level / reference frame that would meet the subsymbolic criteria.³⁵⁰ Thus, predetermining weights that form models at another symbolic level is not the same as predefining symbols for ERMs modeled at the same level.^{351,352} What matters is that this underlying process is outside the symbol set of the reference frame but determines the semantics for that reference frame.³⁵³

Some Considerations for the Criteria Proposed

Before moving on, I want to come back to the claim of information requiring an observer. The less controversial but important claim was that if a system could experience intrinsically, then it would necessarily be grounded in some way. By definition, it would also possess an observer. If that grounding could produce conceptual percepts, I will also claim that it could process information intrinsically.³⁵⁴ However, I also made the claim that the criteria, if met, would mean that an actual system would be grounded. What can we claim about actual systems with regard to intrinsic information processing and possessing an observer? In Chapter 1, I claimed that there is no way to prove a digital model external to us possesses intrinsic emergence and can thus experience. If to observe is to experience intrinsically, then a digital model cannot be claimed to possess an observer and to process information intrinsically from our extrinsic reference frame. In the above, I stated these criteria are sufficient only for determining if a *theoretical* system processes information intrinsically, and that we can only determine the *appropriateness* (i.e., fit) a real system has for this theoretical model. For any *actual* system, we can only determine if it

³⁵⁰ In fact, the vector space is continuous; there is no way even in theory for an external modeler to predefine all the symbols for some large micro and macrostate. For modelers (i.e., those that truly possess semantics), there is no explicit map from the modeler level to the model level. This comes back to the Description Gap; there is a reason ANNs always *learn* features for symbols rather than us being able to determine them *a priori*!

³⁵¹ Consider again that a model formed, when envisioned in its own reference frame has no access to the “symbols” at the level of the ANN weight matrices just as we have no access to our own neural dynamics from our own first-person perspective.

³⁵² Another option to consider would be if grounding can only occur during learning. It would not be an unreasonable assumption if all biological intelligent agents are continuously capable of learning. Nor would it be too bold to suggest that AI also needs to be able to continuously learn in certain ways for it to achieve AGI levels. However, this faces a couple problems. One is that the full training process could be predefined (i.e., same data set, deterministic gradient descent). Thus, it is unclear that it solves the ANN cloning issue. The other issue is that there is no obvious reason that grounding always involves learning other than to address the ANN cloning issue. Plugging holes in a theory is not alone a good justification for adding criteria. Fortunately, I believe the subsymbolic criteria demonstrate that there is no hole to plug.

³⁵³ *Research: Consider in what way delay embedding methods* (Packard, et al. 1980, Takens 1981) *generated for dynamical systems can ground like ANNs.*

³⁵⁴ Whether implicit-only micro and macrostates produce nonconceptual information is unclear. I speculate that such a claim would reach beyond the limits of what can be discussed from the conceptualized notion of Information itself.

acts like the theoretical system with an observer. While this is more meaningful than any strictly behaviorist or materialist alternative, it is still ultimately extrinsic.

However, given how Grounding is defined in this manuscript, if an actual system could somehow be proven based on the criteria, then the claim would be that this system truly does possess the ability to ground symbols. Grounding was defined as the functional aspect of intrinsic emergence, without any claims on intrinsic phenomenological properties. It should be mentioned that, while I have discussed activation of vectors in the context of grounding, there is good reason to believe that, in some way, the entire vector space must be considered. An isolated value on a number line provides no sense of magnitude without surrounding context.³⁵⁵ The same is true for a three-valued vector for saturation, hue, and lightness meant to ground color. Similarly, a higher dimensional vector would not seem any more grounded. At the very least, one must be careful in what specifically would be grounded in a digital model if it met the criteria.

Finally, one might argue that digital models can only ever *simulate* physical processes, including their ability to intrinsically observe. This could be based on the following: 1. We know that at least one physical system possesses an observer (ourselves); 2. Digital models must be simulations since they cannot generate the actual physical causal processes of what they model; and 3. Digital models alone are divorced from the real world (e.g., processing pixel values vs. light energy). Perhaps there is something intrinsically different about certain materials, causal processes, and/or analog phenomena. I will not reject the simulation argument here. However, it is worth considering what to make of this, *as an extrinsic assertion, for intrinsic experience*, given the Percept/Concept-related claims in Chapter 1. Clearly, a digital model of a physical body would only simulate that physical body. However, it is less clear what to make of a claim for an intrinsic observer, since the claim is extrinsic emergent for digital and physical systems alike. This gets to the heart of the Mind-Body Gap and why many remain unconvinced on both sides of the argument of physical vs. digital consciousness. Any extrinsic emergent, and thus conceptualized, system is subject to Eliminative Materialism. We have no reference frame from which an observed physical causal or a digital system can explain how intrinsic phenomena would occur for them.³⁵⁶ No matter how natural they seem, requirements such as having inputs directly from the physical world, actual neurons, physical causes, the ability to internally simulate/model/predict, self-awareness, etc. do not seem to help. While there is no clear reason to reject this simulation argument, it is worth future consideration for what that means specifically for intrinsic emergence with respect to the fundamental role of Reference Frames.

Regardless, given that extrinsic emergence does not favor any particular type of actual system, I am compelled to claim that, for *any* actual system, the grounding criteria provide only

³⁵⁵ *Research: It is worth considering if/how Grounding involves emergent dynamics associated with how vectors are constrained by their dimensions or by other vectors. Grounding (and perhaps Observation and Intrinsic information) might only make sense as Processes. Thus, certain dynamics may be required that are not captured by instantaneous activations associated with vectors, and these may be very limited even in current AI models that seem to meet the criteria proposed in the main text.*

³⁵⁶ While we know of at least one physical system that possess an observer, this is an association with no explanatory power.

empirical evidence. This is consistent with what one can really claim for other humans. I thus take the more general view that we could only ever say *any* actual system “acts as if” it has an observer and processes information intrinsically. Note that “acts as if” is all we need for AGI (as long as it acts like humans). The functional aspects of grounding, observing, and intrinsic information processing are covered. All that is left out are the phenomenological aspects of intrinsic emergence. For understanding the Mind more generally, the claims at least shed light on the fundamental role of Reference Frames. Furthermore, **the subsymbolic criteria proposed above are sufficient to claim the following: While they cannot be proven from our extrinsic emergent perspective to be anything but acting as if they process information intrinsically, semantics/grounding is functionally performed by systems that meet these criteria.**³⁵⁷

Since the theoretical model for which we take the reference frame does possess grounding, an observer, and intrinsic information processing, there is another approach I will consider for assessing actual systems. This approach assumes a given digital model can only ever approximate the theoretical model, which we will assume is a Conceptual Space. However, it then makes the more speculative claim that the digital model is equivalent to this theoretical model “in the limit”. What is meant by “in the limit” will also be described.

Percepts/Concepts and Semantics/Information

The claims in the above section reflect my earlier claims that Grounding and Semantics are equivalent. Interestingly, these criteria are another way of saying that Semantics/Grounding would be Percept-level measures (and thus Percepts must be subsymbolic at their own level / reference frame), while Shannon-like Information is measured as Concepts (and thus Concepts always involve Symbols, and a macrostate is always about something). In fact, there really seems to be an equivalence between Semantics and Percepts (or perhaps as the functional equivalent of Perceptual experiences), as well as Information and Concepts. Any phenomena outside a symbol set from a given reference frame are Perceptual. If Semantics is equivalent to Percepts, and Information to Concepts (or more accurately, the explicit micro and macrostates against which information is processed are the conceptual property of certain percepts), then this brings with it all the equivalencies discussed in earlier chapters, such as first- vs. third-person perspectives, and subjective vs. eliminative properties. The use of Reference Frames in the subsymbolic criteria above is nothing more than an attempt to consider models from their theoretical first-person perspective. As I have already said, something symbolized is

³⁵⁷ *Research: If Semantic Processes are functional, then is there a way to compare Extrinsic Information and (acting as if possessing) Intrinsic Information gain for AI models?*

conceptualized and has intentionality since a symbol is about something.^{358,359} Likewise, a concept could be said to be a symbol. Symbols emerge from the subsymbolic Semantic Processes. Shannon-like Information is content-less unless grounded in semantics (model formation), just as Concepts are special cases of percepts that have properties of being about percepts.³⁶⁰ Concepts alone are eliminable, and symbols alone are empty.

This comes back to the models vs. modelers claims as well. ERMs are not modelers because they are modeled at the conceptual level. They are only third-person in that they are about something just as any text in a novel etc. is about what it references.^{361,362,363} And just like text, they do not have the processes required to perform the referencing themselves. They can only evoke the references in external modelers. They have no perspective of their own. Thus, ERMs represent our concepts without any of the intrinsic perceptual properties that we as modelers

³⁵⁸ These claims for Semantics may explain why consciousness is so hard to grasp: Consciousness is the phenomenological properties of grounding and therefore can only be understood at the levels within which symbols would emerge. The moment we conceptualize this conscious process, we are moving to a different level, where the conscious process itself has been symbolized. That symbol is at a different level from the one that would be required to truly “explain” conscious processes at work forming models. As a symbol, it is subject to eliminative materialism, and yet that symbol is accompanied by the semantic grounding it generates when we receive it, and this creates the issue that something seems to remain outside the eliminable symbols.

³⁵⁹ *Research: Much more speculatively, If Semantics is necessarily subsymbolic and tied to percepts, while Shannon Information is necessarily symbolic and tied to concepts, then consider if this could in any way relate to quantum-level wave function collapse. Does measurement conceptualize what are “naturally” non-conceptual phenomena? I am not sure what this would mean. However, I have no doubt that this Percept vs. Concept and Subsymbolic vs. Symbolic and Semantics (Meaning) vs. Shannon Information is core to how we understand our reality (even more so than Information itself), so it is worth considering potential implications even in fundamental areas of science.*

³⁶⁰ The use of the term ‘Conceptual’ in ‘Conceptual Spaces’ clearly conveys that they can represent cognitive processes, such as inference over concepts. As mentioned in Chapter 3, given the way I describe Percepts and Concepts, I see Conceptual Spaces as capturing not just Concepts but Percepts in general (when taking their reference frame). Given that Concepts are types of Percepts, the term “Perceptual Spaces’ would convey that they capture a larger range of cognitive processes without minimizing its use for high level inference. ‘Concepts’ only makes sense if referencing how we as external observers ground observed activations in the Conceptual Space. I also emphasize the issues with assuming concepts are fundamental to inference, and that percepts are the appropriate form for modeling grounded AI. I point to Conceptual Spaces as being capable of grounding, which a model starting with Concepts cannot achieve. There should be no confusion as long as readers understand that I assume Conceptual Spaces naturally represent Percepts, which can include but are not limited to Concepts.

³⁶¹ By this, I do not mean that ERMs are intrinsically Concepts / Third-Person, which requires a Perceptual / First-Person base. I mean that this is what they represent to an external modeler. What they symbolize is not symbols to itself.

³⁶² Of course, text can be written in first-, second-, or third-person, but this is a derived property that must be evoked in external modelers to be understood as such. The same goes for ERMs. As concepts without grounding, even their intentional nature is only experienced within grounded observers.

³⁶³ Note that my point is not that symbols do not serve a function, including in reasoning processes. We have concepts because our world exhibits regularities from which we can effectively abstract concepts for them. This allows us to compress signals, saving energy and time, as long as it is effective for our reaction propensities to be the same for all the categorical or analogical entities covered by a given abstraction. What is at issue is assumptions about how symbols can be modeled without the grounding that drives their functions. The function of symbols is severely diminished without the subsymbolic base that generates them. Conceptual Spaces and associated vector types (e.g., force vectors) provide ways to model those symbols, such as by using similarities and vector operations across concepts as analogies for new processes and force vectors.

experience with concepts. These perceptual processes are Semantic Processes. If this equivalence holds between Percepts/Concepts and Semantics/Symbols, then, **just as concepts are a property of a subset of percepts, so, too, symbols would be a property of some subset of a subsymbolic base.** What this does is convert a vague complaint about ERM's lacking the ability to model implicit properties required to ground symbols / be emergent into a more specific one where ERM's lack the ability themselves to map symbols to referents, no matter how many relations and concepts they model.

Thus, the claim is that **Percepts are equivalent to the subsymbolic base that generates Semantics. Percepts just "are", making them fundamentally subsymbolic in and of themselves. This is also why I claim they have no intentionality, which must involve symbols that "point to" something else that they are symbolizing.** Intentionality is a derived property of Concepts.

This also suggests another possibility: **All explicit Information measures are the extrinsic side of Semantic Processes, which are intrinsic.** This does not contradict the claim that concepts are properties of certain percepts, since concepts are still percepts (they are percepts about percepts). Percepts are more fundamental, and likewise, only some Semantic Processes would possess Information processes as extrinsic versions of other Semantic Processes.

How could we measure Non-Conceptual vs. Conceptual processes for an AI model? One option could be to require that the symbols included in macrostates formed by the model are not the whole of what the model "processed", and that those symbol sets must be formed by these other processes. These are behaviors that should be observable externally by assessing if the AI forms percepts for inputs that are not about anything from the theoretical perspective of the AI model and if it can build models from this medium of nonconceptual (nonintentional) processes (i.e., build processes with intentionality, in that they map to concepts in some world/domain, from something fundamentally non-intentional). This can be measured by considering if the end-to-end features/variables for the AI in question are reducible to the level and set of concepts modeled (i.e., that make up the AI's macrostate indices). ERM's would be reducible in this way (as long as the predefined mapping done by the external humans is accounted for). ANNs, on the other hand, would not; they have at least one step where the original labeled concepts cannot be reduced to the learned distributed representational formats within the ANN's weight matrices. Indeed, to compare any explicit concepts in a Semantic Vector Space requires forming them from the space itself, from which they fundamentally map to a vector of values across various quality dimensions and distances to other percepts/concepts. Concepts cannot be explicitly mapped at the same level as their distributed representations. Therefore, one of these can be considered subsymbolic relative to the other, depending on which reference frame we consider. Thus, if one reference frame for an AI model depends on properties outside the symbols within that reference frame, then those left out of it can in theory ground the symbols within that reference frame. This is what can be measured by an external observer, and what we might assume would remain the case for the AI were there anything for it to be like. That is, **for a given reference frame (assessing them ourselves externally) for AI models, if properties about data received or modeled are not symbols, and they are used to determine what symbols map to, then we may conclude that, if it were possible for the AI model itself to**

take that reference frame, then it would also leave processes outside its symbol set that are involved in determining the symbols. This also means that we can say that, if the AI were to process anything, it would have intrinsic non-conceptual perceptual processes.³⁶⁴ The grounding associated with this subsymbolic base just “is”, just as a percept just “is”. Any experience associated with the grounding/percept would also be simply what it is, which is not eliminable without transforming it into a concept.

Related Grounding Theories

(Harnad 1990) first proposed the Symbol Grounding Problem, as well as criteria for being able to ground symbols. Harnad refers to a solution for the grounding problem involving the use of subsymbolic processes to “*pick out, from their proximal sensory projections, the distal object categories to which the elementary symbols refer*”. The idea of picking out categories bears some likeness to my claim that the subsymbolic processes must form the micro and macrostates. The value I see in describing the role of the subsymbolic processes in terms of micro/macrostate formation (or index construction) is that it specifies something for an agent/model to incorporate, including coupling with Shannon-like Information processes that are measurable over models. This hopefully helps to implement a generalizable process for determining the meaning of message symbols.

My claims are also very much related to (Haken and Portugali 2014),³⁶⁵ who propose that Semantic Information selects the indices for Shannon Entropy measures. Again, this means that Semantics selects the variables assigned to i in the Shannon Entropy equation (Equation 4).

$$H(X) = - \sum_i P(x_i) \log_b P(x_i) \quad (4)$$

The only obvious difference from my claims is that I call this index selection process Semantic Processes, rather than categorizing it as an Information type for reasons already described.

The claims for Semantic Processes can also be related to research in Gestalt Principles (Todorovic 2008). **The grounding of Gestalt properties given data elements relates to how the combination of those elements evoke changes in perceived geometric forms. A theory of Semantics would thus relate to these geometric forms, and how they are evoked by messages. A model formed may abstract over multiple elements, creating an emergent symbol. A different model could have been formed over a different set of elements, including one symbol per each element. This is equivalent to what I mean by model (i.e., micro and macrostate) formation and is thus equivalent to Semantic properties.** As I will discuss later, what forms are evoked in us is constrained by Effectiveness, often over our evolutionary history. Since each modeling decision can be represented as a geometry in a semantic space, Conceptual Spaces offers a way to measure how form relates to Semantics. **From such forms,**

³⁶⁴ This includes grounding that underlies any concepts, since concepts are always properties of certain percepts.

³⁶⁵ In fact, I did not realize that my claims for Semantic Processes could be viewed as the symbol index in the Shannon Entropy equation until I read this book. My use of this description is owed to (Haken and Portugali 2014) and help make the associated claims more tangible.

and changes across them, we can understand how Gestalt phenomena are generated, which cannot be understood through the discrete combination of symbols (e.g., words) alone (i.e., without their implicit properties which are represented geometrically in semantic space).³⁶⁶ Gestalt properties cannot be separated from Semantics, which we have already described as emergent heuristics.^{367,368} It is through a subsymbolic base that Gestalt processes are possible. Data impact the subsymbolic base, which evokes forms that are mutually constrained by the data received and the subsymbolic medium. These forms are what map to symbols. To measure Semantics, if it is possible to do so, we therefore must consider how these geometric forms could relate to Interpretation. I relate a notional and very basic geometric relation to measuring Grounding in a later section, but this is an opportunity for considerable study.³⁶⁹

Twenty Questions

As mentioned, this is a difficult subject to explain, given the necessity of using symbols for any explanation. The very conceptualization of a process forces it instead into a symbol about that process that is open for Shannon-like modeling. The grounding measure is, however, distinct, and it is what really matters for bringing Information internal to a model. Without it, AI will always require humans to do the mapping. In this section, I use the game of ‘Twenty Questions’ to help understand Semantic Processes and Conceptual Spaces.³⁷⁰

Let us say we know the Twenty Questions topic is Animals. The Shannon macrostate is the set of possible symbols that could match the Twenty Questions solution, which is the microstate. Each question reduces, to varying degrees, the macrostate to a smaller ensemble of animals that could match the solution. This reduction is measured as Shannon Information gain, and the uncertainty about the possible microstate given the updated macrostate is measured as Shannon Entropy.

If an AI model has every known animal and their properties all predefined as “symbols”, and a rule is added to ensure the model understands how the two possible messages ‘yes’ and ‘no’ must update the model, then any solution can in theory be found and the entire process can be

³⁶⁶ *Research: Consider how Conceptual Spaces, and their Semantic Vector Space approximations, might facilitate finding regularities in emergent properties. That is, use the embeddings developed to discover possible regularities in concepts given their association with known lower-level concepts. They should provide an empirical complement to theories and/or common sense about how emergent concepts would be expected to relate. What patterns remain to be discovered for emergent phenomena through analysis of them in the context of their embeddings?*

³⁶⁷ *Research: Consider if knowledge in Conceptual Spaces is less reducible than other methods because each knowledge element is only understood by its geometric relationships to all other points in the continuous space. This makes sense if simply enumerating the given knowledge element by its value along each concept dimension does not fully capture the given element’s context. That is, there are also other properties such as density related to all other points, etc.*

³⁶⁸ *Research: The relation here also suggests that Gestalt principles should go well-beyond the standard visual-level principles. Gestalt effects should underlie cognitive phenomena all the way “up” to higher reasoning. Consider research to discover how the Gestalt principles can be discovered across all cognitive “levels” (or rediscover what has already been proposed by Gestalt researchers).*

³⁶⁹ *Research: Develop these research study concepts, including leveraging and integrating the considerable amount of existing research in both Gestalt and Conceptual Space literature.*

³⁷⁰ Indeed, writing this section forced me to change some initial claims on Semantics vs. Shannon Processes.

measured using Shannon Information. Expected Shannon Information gain can be measured based on the proportion of each property across remaining animals, and each question that is expected to maximize information gain could be asked automatically by an AI model. Ignoring all the practical challenges (e.g., tractability, engineer labor), what is described here is an ERM, where the macrostate is analogous to a belief about the world (i.e., microstate), with the uncertainty of beliefs measurable and updatable using Shannon Entropy. I claim there is no Semantics in such a model.^{371,372,373} Semantics can be evoked in external observers of the model, but the model itself contains no Semantics. A human had to determine the symbols that make up the macrostate (i.e., the set of possible animals) and the rules by which to update it. A question, then, is if any grounding was required by the human to create the model.

First, we should note that the above does not describe how humans model the problem. While a computer model may be able to store and structure most of the animal properties required to play this game well, it seems unlikely that humans would have such a structure. Right now is the first time I have ever considered whether a Cooper's hawk can fly, but I know the answer. It is unlikely that I activated for the first time some explicit symbol in my brain for the flight attribute of the Cooper's hawk. It is more likely that some underlying subsymbolic process was used to form any symbols for Cooper's hawk and to infer whether they can fly. The number of potential properties is essentially open ended, hence the need for a generative process that enables generalization through implicit analogical processes. Of course, such a generative process cannot provide perfect knowledge. I do not know which set of animals have between 8 and 19 teeth or spin in circles more than once a day at an angular velocity of $>5\text{mm/s}$. However, I understand the properties and know that they could be asked as valid questions. Even with my limited knowledge, there is simply too much to model everything explicitly without making use of some underlying grounding. Perhaps more importantly, it is not even clear that everything that matters can be symbolized. At best, ERMs face the practical problem of how essentially boundless information could be modeled.

Conceptual Spaces are much more plausible models for Knowledge. They can in theory represent all these properties and animals without having to explicitly represent them. Recall that a Conceptual Space consists of quality dimensions, where vectors represent object instances. Specific animals may represent subspaces in a vector space with a hypervolume proportional to the value ranges across some set of those qualities. Although this an oversimplification, each quality dimension can be thought of as being mapped to the attributes

³⁷¹ Modeling properties such as adding conditional relationships, including hierarchies to share properties common to different animals (e.g., a set of reptiles, birds, etc.) would not change the fact that there are no semantics.

³⁷² Also recall my claim that, since there are no semantics, the symbols in this model are only symbols to external observers. There is nothing internal to the model itself to interpret distinctions as symbols.

³⁷³ Note also that the uncertainty measured for a model will depend on the indices selected by semantic processes. Thus, the performance ceiling of probabilistic ERMs, such as Bayesian models, depends on the semantics the modeler applies to form the model. Grounding thus enables selection among possible indices based on some implicit criteria to optimize, whereas handcrafted model indices are fixed, at least until manual edits are made. This is another major limitation, and it at least partially explains why humans are better decision makers than such AI models, except for limited problem types.

that could be used in questions. Any given location or convex region in the Conceptual Space could represent some concept that the owner of the Conceptual Space has never activated before, creating a concept that is both implicitly ‘known’ and yet never previously explicitly considered. The fact that this is a metric space is critical for grounding. Saying a given space can ground vector positions means that its geometry, including distance to other locations, is meaningful. Dog breeds will tend to be near each other and also near wolves along dimensions for size, speed, vocalization properties, etc. Along some (likely fewer) dimensions, however, dogs will be nearer to cats than wolves, such as in living space.^{374,375,376,377,378}

What is happening with the Conceptual Space during Twenty Questions? What is to be measured as Grounding vs. Shannon Processes? At least part of the answer is the game itself. When the game is constructed symbolically, the symbols received (yes or no) are of the same symbol set as the question asked because they relate to the same macrostate, as long as someone has implemented their application rules so that the logical operations of the two possible answers appropriately update the macrostate. Humans understand the concept of the game and that the scope is animals, such that yes/no messages received will be applied to reduce the macrostate of animals potentially matching the microstate. They can thus in theory construct a computer model of the game (e.g., answering human questions) or for playing it against a human (e.g., asking some limited scope of questions to the human to find the solution). That is, an ERM is capable of modeling ‘Twenty Questions’, although limited to the scope that different possible categories and their explicit properties have been constructed for it. Although the game is technically unbounded, one can imagine that upfront work mapping animals and their properties to a macrostate can handle many questions. Once the categories and their explicit properties are in place, Shannon Information is all that is needed.³⁷⁹ For Shannon Information, each answer to a question is a message that reduces the macrostate (possible answers) towards the specific microstate (actual answer). In this sense, Shannon Information increases Specificity of the macrostate with respect to the microstate, and some messages increase specificity more than others.

³⁷⁴ Geometric properties also enable locating appropriate referents for symbols in the message and the larger set. Consider analogies: Learning that dice are like coins but with six sides enables translating much of the grounding already provided by coins, which is required to extract information out of dice events (e.g., rolling a 4), such as modeling number of dots as magnitudes or that the sides should be equiprobable.

³⁷⁵ I only describe Conceptual and Semantic Vector Spaces enough to justify my claims. In fact, many of the ways I apply them are oversimplified. Describing the wealth of research in these areas is both out of scope for this manuscript and redundant to existing literature. **I strongly recommend that any reader interested in understanding Conceptual Spaces read through the research by Gärdenfors, his collaborators, and others who have developed insights and theories in this area. I also suggest reading articles to understand Neural Embedding methods and how they are trained to model Semantics to perform prediction and inference.**

³⁷⁶ **Research:** (Gärdenfors 2014) *provides many examples of how vectors relate to semantic properties (e.g., concepts as force vectors). Account for how these can all be measured in AI with semantic processes.*

³⁷⁷ **Research:** *Implement all the research and insights on Conceptual Spaces to use for Semantic Vector Spaces.*

³⁷⁸ Note: There are many possible distance metrics for metric spaces. I make no claims as to which are best for AGI.

³⁷⁹ Keep in mind that the Information processed would be extrinsically projected, since the ERM is ungrounded.

Is Grounding necessary? Besides the implausibility of their being some ERM in the minds of humans, we clearly ground the messages received. The receiver must apply each symbol ('yes' or 'no') in the message to the model (macrostate) for the message symbols to reduce it. Selecting the macrostate to reduce (or one could say, "how" to reduce the macrostate) requires a semantic interpretation of the message symbols. This grounding is what I have claimed are the subsymbolic Semantic Processes that decode symbols received to form the macrostate to update, which modifies the subsymbolic space and can encode new symbols to ask new questions.

Since a Conceptual Space consists of a continuous hypervolume, it cannot be fully mapped to any finite set of symbols. The Conceptual Space can emerge symbols given messages, but this does not eliminate the subsymbolic role in that emergence process.³⁸⁰ In fact, real Conceptual Spaces should be able to take discrete message elements and emerge something different than the sum of the parts. Operations evoke semantics associated with certain locations in Conceptual Space. Locations or associated geometric forms should be associated with emergent experience. For example, summing or averaging produces new vectors or subspaces and therefore changes across some set of dimensions that determine qualities. Thus, vector operations create different semantics that could be experienced as different qualities. This is made possible by the geometric properties that a Conceptual Space provides by being a metric space, which allows for meaningful geometric relationships. As a simple example, an operation to change the size of a subspace could represent a more abstract category (e.g., from cat to mammal) with implicit semantic qualities not captured by the explicit properties in any of the individual members. One can envision the effectiveness of these operations in how it can ground concepts given messages.^{381,382}

Thus, Twenty Questions requires interpretation of symbols as a model formation process. Grounding of the symbols must be subsymbolic and seems to require something like a Conceptual Space to learn the grounding for concepts. Like the Chinese Room Argument, a human can in theory construct an explicit model that can measure expected Shannon Information gain through counting properties but doing so requires that the human had already grounded concepts for the symbols they generate into the computer model. It is only by including the human modeler as part of the model that one properly accounts for it.³⁸³

³⁸⁰ Similarly, ANNs map features to distributed processes and bear no *explicit* relation to the semantics we would impose externally on a set of labels for any symbols of the micro and macrostates they form.

³⁸¹ This process may describe how Mental Chunking (Chunking_(psychology) n.d.) occurs and is experienced.

³⁸² *Research: Increasing research across Cognitive/Psychology/Neuroscience and Conceptual/Semantic Vector Space modeling is important. For example, how do different interpretations for increasingly ambiguous messages occur (e.g., what triggers selecting one over another? What about when different interpretations involve different vector operation types?)*

³⁸³ I should note that this example for Twenty Questions is much cleaner than what would happen in practice, at least for humans. Some symbols may not be interpretable if the answer is an animal I know nothing about. This is a case where the symbol cannot be sufficiently grounding in the Semantic Vector Space to update the general macrostate applied to the game. Also, during questioning, there may be some properties the modeler has not considered. Humans are not always capable of maximizing expected information gain, especially children who tend

Any explanation of the Semantic Processes is unfortunately difficult, as the symbols used convert the explanation to Concepts rather than non-conceptual Percepts. I can however suggest that the reader play Twenty Questions and reflect on the experienced process. They can of course choose to explicitly approximate expected Shannon Information gain, but to do so requires conceptualizing the otherwise intrinsic model. The player becomes a third-person observer to the model they assess. Using a computer aid to measure expected information gain could be helpful in some cases, but of course a human does not have to play this way. Moreover, the player should be able to note that they still need some implicit process to construct this explicit model. Grounding is necessarily involved in forming a model at any level.

Subsymbolic in the Limit

The subsymbolic and intrinsic information criteria I provided earlier were a theoretical construct. The challenge of determining whether a real system meets those criteria is shifted to one that cannot be proven but can be claimed to behave in a manner that is consistent with meeting the criteria if they did. This cannot prove anything regarding actual intrinsic processes. As stated in the Chapter 1, this is a particularly clear challenge for digital systems due to their discrete nature,³⁸⁴ since we know and have access to their underlying mechanisms (i.e., bits). I have spent considerable time discussing ANNs, and in particular, the Semantic Vector Spaces they underlie, and claiming they seem to fit the criteria to possess subsymbolic and semantic processes. Yet, these types of AI are digitally modeled, as is most AI. All we can really claim about such models is that they approximate, or perhaps simulate, the phenomena they model. Because extrinsic emergence involves conceptualizing phenomena, those phenomena are necessarily represented in a symbolic form. This means that there is always some reference frame that can conceptually approximate digitally implemented Semantic Processes with arbitrary levels of fidelity.³⁸⁵ That is, a modeler can be conceptualized as a model from the third-person perspective.

However, an idealized Conceptual Space is not symbolic because it is a continuous metric vector space.^{386,387} As a theoretical construct, there is no symbolic model that can conceptualize it. Thus, they can only be approximated in digital systems. In what follows, I leverage an approximation process for Conceptual Spaces to propose a stronger claim about Semantic Vector Spaces' potential for intrinsic and subsymbolic properties: **Even though any**

to ask very specific questions (e.g., Question 1: "Is it an ostrich?"). None of the nuances I can think of that affect real game playing weaken the actual arguments made in this section.

³⁸⁴ At least, if assuming a Materialist Reality.

³⁸⁵ This does not contradict the subsymbolic criteria, since it requires a different reference frame, but it is why those criteria are unprovable for any actual systems.

³⁸⁶ *Research: Could this "in the limit" claim also then apply to intrinsic emergence, and therefore help explain consciousness? That is, is it only mimicked when approximated in the way explained above in any digital system as a Shannon Process but can be said to be conscious in the idealized limit?*

³⁸⁷ *Research: Determine if it can be proven that Semantic Vector Space representations can get better than or as close as any other possible representation type applied on digital systems, since the method models continuous space, such that the limit really is at the resolution of the computer (and can be even better with numerical solutions applied, at the expense of processing time).*

extrinsic emergent processes can be eliminated to their mechanisms, and digital systems consist of discrete processes without room for intrinsic emergence at their own level,³⁸⁸ it is possible for digital models, under certain conditions, to approximate the semantic processes arbitrarily well. When these conditions hold, we might speculate that, *in the limit*, the approximations achieve Semantics and Intrinsic Information processing.

Although I will not fully determine these conditions, given that Conceptual Spaces have a strong claim to subsymbolic processes that meet grounding requirements, Semantic Vector Spaces may be the type of approximation that can achieve Semantics in this limit. They clearly approximate a metric space. Since Semantic Vector Space are digitally modeled, convexity is not truly represented, but it can be approximated arbitrarily well with sufficient network complexity and the right type and amount of data. Therefore, my hypothesis is that **Semantic Vector Spaces, in the limit achieve true semantics. I tentatively claim that the limit here refers to the upper numerical bound for the metric space fidelity.**³⁸⁹ While I will not attempt a proof, this limit-based argument may apply uniquely to Semantic Vector Spaces and possibly other distributed architectures, where symbols are replaced by vectors of numerical values. The approximation of a Conceptual Space (with its continuous and metric space properties, among possibly other key properties) may be uniquely suited for this claim.^{390,391}

Note that I am not claiming that any actual Conceptual Spaces might be digital. The limit for digital models is unreachable. I am also not claiming that Semantic Vector Spaces in their current forms are arbitrarily good approximations of human conceptual spaces, and certainly not in how they are trained. There are limits to learning passively from data, let alone purely text or image data (compared to how natural minds learn and have evolved). Moreover, ANNs are almost nothing like real neuronal networks. It may even turn out that some specific embedding method must be applied that remains undiscovered. This brings up an important general point about how AGI is tackled today. Ultimately, discovering new methods for

³⁸⁸ I have resisted making many metaphysical claims in this manuscript. However, it is too obvious not to mention that one possibility for understanding consciousness is to claim that qualia (i.e., intrinsic emergence) are simply explained as relative with respect to reference frames. That is, a condition for qualia existing is that the reference frame from which they exist differs from the one in which they are symbolized. Reference Frames thus take on a fundamental ontological status (at least as fundamental as qualia). As I have mentioned, this explains why an explanation for qualia always eludes itself due to the necessary conceptualization (and why qualia are yet so familiar even without explanation). However, this is not very satisfying as an actual *solution* for understanding qualia.

³⁸⁹ *Research: Consider if there is a relationship between Zeno's Dichotomy Paradox (Zeno's paradoxes n.d.), Searle's Chinese Room Argument, the above claims that digital models can never be conscious, and that Conceptual / Semantic Vector Spaces may be able to approximate intrinsic processes in the limit. The specific relationship to explore is the fact that digital processes in a sense avoid Zeno's Paradox by moving straight from point A to be point B since bits are discrete. Is there some fundamental relationship between Zeno's Paradox and continuous processes, thus tying intrinsic emergence / consciousness / qualia to this continuity in the environment (and/or mind)? Is what is not yet explainable for qualia the very same as what is not yet explainable about how we actually are able to move in the world? Should any solution to Zeno's Dichotomy Paradox also solve how qualia / intrinsic emergence are possible?*

³⁹⁰ *Research: Determine how this can be proved or disproved.*

³⁹¹ ***Research: Determine an architecture and multimodal training design to research and develop Semantic Vector Spaces in analog devices (e.g., neuromorphic hardware)***

representations is much more important than trying to improve or integrate existing methods that do not fundamentally resolve how to represent the semantic space itself.

All that “in the limit” means is that a given model that meets the grounding criteria provided will be the same as *some* theoretical model that would ground symbols, have an observer, and possess intrinsic information processing. There is some possible theoretical model that a given actual model becomes in the limit, as long as the actual model meets the grounding criteria provided in this chapter. It makes no claim about the properties of the grounding that would be achieved for any particular model.

This “in the limit” approach is nonexclusive to the earlier criteria put forward for something to be subsymbolic. It aims to show that the problem of mechanism-level sufficiency would disappear “in the limit” for digital implementations of Conceptual and Semantic Vector Spaces as the metric space they construct increases in fidelity with respect to some theoretical Conceptual Space.^{392,393}

Of course, this “in the limit” claim is attempting to address a problem that has some relation to the Mind-Body problem, which is how mechanisms can give rise to emergent processes intrinsically. While I have focused on digital systems, the claims made seem equally suitable for physical mechanisms. However, the criteria here only provide a theory for *when* rather than *how*. Still, the possibility of consciousness in AI and computers is hotly debated and is already considered an important topic, and it therefore seems worth considering if there are any useful implications. I have not fully considered if/how the claims here could be used as criteria but, if they are correct, then they may offer some insight into how we go about making such determinations.

Finally, I want to address what these criteria might say about grounding in The Chinese Room argument. I am not addressing whether or not the person in the room intrinsically experiences the symbols they receive. The person in the Chinese Room, as it is laid out in the original scenario, would also have very limited grounding. This is not because they do not have the potential for grounding or the ability to intrinsically process information. Humans meet the criteria in this chapter. The problem is the forms of data available to the person. The symbols are initially meaningless scribbles. All the person observes is how some scribbles relate to other

³⁹² Essentially, “in the limit” refers to properties for certain model types that can be said to approximate a continuous metric space as its fidelity. Such an idealized space cannot be considered to be made up of discrete bits and thus is necessarily outside the symbol set received, as well as any symbols formed by its non-bit subsymbolic processes. This idealized model reached at the limit is equivalent to a Conceptual Space and Semantic Vector Spaces seem to be valid approximations.

³⁹³ Besides Fidelity, I have only considered Span or Coverage for this limit. However, humans do not have infinite Span or Coverage, so these do not seem appropriate. Note that fidelity here is not a claim about the fidelity of the perceptual properties (e.g., the information capacity of neurons for messages) but for the analog biochemical substrates that Semantic Vector Spaces would be approximating. Clearly, much more consideration is required, such as how far down the mechanism level is required for the substrates, what it would mean to reach their “fidelity”, all while avoiding conflating this substrate with its conceptual form. In short, this is a vague idea in need of much more thorough consideration. Still, the subsymbolic nature of theoretical Conceptual Spaces are intriguing enough to raise the idea of if/how it might be approximated arbitrarily well. And of course, Semantic Vector Space models, which seem to already be approximating them to some degree, seem like a good place look for insights.

scribbles. The human can only learn to ground symbols from these scribbles. This is grounding, but it is not very rich and certainly not a grounding at the level of what the symbols mean to Chinese speakers, including whoever authored the book they are using. Grounding at this level requires other types of sensory data to implicitly connect to the symbols, and so will AGI. However, AGI also needs many of the other things that the person already has in virtue of being a human, including many of the things discussed in other chapters of this manuscript.

I have attempted in the above sections to thoroughly address what can and cannot be claimed for actual systems with respect to Grounding and related processes. In the sections that follow, I will assume the reader is aware of what can be claimed and will drop ‘approximation’ for grounding discussions to improve readability.

Measuring Grounding

How can Grounding be measured? Is it a magnitude? Is it a vector of properties? Is it only qualitatively measurable? A scalar value clearly would not measure anything about the qualitative properties of any Grounding. In a Conceptual Space, it is unclear that there is anything less than the vector itself that describes its qualitative properties.³⁹⁴ Attempting to measure semantics with a single value may be like trying to measure intelligence or body shape with a single value. On the other hand, there does seem to be more and less grounding. It is plausible that two agents with less experience with a given symbol will not only have different qualities for the grounding but different magnitudes as well.^{395,396} If so, then there is considerable value in having a way to determine how much different systems ground the same symbols received to augment black-box measures based on outputs. In this section, I propose some options for measuring the magnitude for grounding.

Here are some starting requirements for a Grounding measure:

³⁹⁴ A similar distinction in measuring distances in vector space that could be considered to help resolve whether grounding and/or Semantic Information must be vector-valued, such as the scalar distance between a starting and ending location or region in semantic space evoked by a given message vs. the vector value that describes the delta between a starting and ending location or region in semantic space evoked by a given message. However, other measures can only be understood as vector values, such as the resultant vector itself activated by a message.

³⁹⁵ *Research: Consider if natural selection adapted semantics over a set of identifiable factors that can be applied for multi-objective optimization, or alternatively certain “knobs” that can be dialed up or down to modify the level of grounding for that particular property. Other options to consider would be surrogate measures “selected” for their tendency to produce effective results with respect to an evolutionary fitness function. This could include associations with measures based on Semantic Information theories, such as learning truths about the world, disambiguating semantic interpretations, or any other measures that would tend to improve coordination and decision-making.*

³⁹⁶ *Research: Consider how to apply embedding spaces for measuring grounding over sequences of vectors rather than just as static vector outputs. Thinking is a dynamical process. When we are listening to a story or reading a paragraph, how is it determined when to start and stop grounding it in chunks? That is, some sequences of vectors would be expected to be grounded together to achieve a new single vector (e.g., black cat, little league). For many sentences, the initial words influence the grounding at the end of the sentence. How does the conditional grounding or path dependence for vector operations “decay” over sequence length? The process likely involves some recursion, which means that our brains (subsymbolically) determine, when receiving a sequence, to use memory of the sequence to reprocess it given more recent elements. These could be assessed with a Shannon-like measure, such as prediction error. How would these determinations be made?*

- If grounding can be associated with a magnitude, then it should be associated with something like the richness of experience or amount of context evoked for a given a message.³⁹⁷ Such a measure should be smaller for an untrained or less trained Semantic Vector Space (or one that has received little data) relative to a well-trained model given the same messages received.
- The measure should not rely on some external reference (i.e., some counterfactual set of possible interpretations) for reasons described elsewhere in this chapter.
- A grounding over symbol sequences should, in some cases, differ from the sum of the grounding of each symbol in the sequence. This captures the semantic phenomenon of mental Chunking (Chunking_(psychology) n.d.).³⁹⁸
- By forming a micro and macrostate, a grounding measure should be consistent with Semantics' role, including actualizing symbolic-level inference as measured using Shannon, Semantic, and other Information Theories.

I will focus on Semantic Vector Spaces, embedded using ANNs, to propose a potential Grounding measure. The only way I can think to measure magnitude of grounding is conceptually (i.e., symbolically). Thus, we can use Information-based measures (e.g., Minimal Description Length and Shannon Entropy) but for the level that is subsymbolic to the symbols they form in the Conceptual / Semantic Vector Space (e.g., neural / vector-level properties). As discussed, they can be used to measure the grounding as long as what symbols are being used in the measure are not from the same set of symbols that are encoded and decoded by the grounding.

A common method for training a Semantic Vector Space is to predict cooccurrences³⁹⁹ of, say, words, and adjust edge weights based on differences from the predicted and actual cooccurrences. Each node (i.e., neuron) in the ANN represents a dimension in the Semantic Vector Space, and each word is a vector whose coordinate is made up of its weight values across the nodes/dimensions. Before an ANN is trained, assuming random initialization, the ANN should have essentially no semantic understanding of a received word. The Semantic Process forms the semantic space vector for the input word, which determined what word is output as a prediction about a neighboring word(s). The backpropagated weight changes calculated from the prediction error modifies the (semantic) grounding of the semantic vector space. This is the subsymbolic semantics underlying the updated Knowledge System. This process is repeated many times using available training data.⁴⁰⁰

³⁹⁷ This does not contradict the earlier claim that grounding can occur at any level, even for messages in a foreign language. Less context is available to ground such an experience relative to the experience of, say, listening to one's favorite song.

³⁹⁸ Chunking is discussed later in this chapter as a product of Semantic Compression.

³⁹⁹ Semantic Vector Spaces should be trainable in other ways than cooccurrences, but this is currently a common approach and so is used as the example.

⁴⁰⁰ *Research: Consider how certain types of inference processes could be integrated with Conceptual Spaces. How is pattern finding understood in terms of conceptual space properties and processes? What about Logic? Math? Intuitive Physics? Consider how Force Dynamics concepts (Talmy 2000) can be integrated into Conceptual Spaces. Force Dynamics constructed from within Conceptual Spaces seems like a potential solution for AI learning intuitive*

If Semantic Information measures reduction in possible interpretations (either about what was sent or what it asserts about some domain), then I believe what we need to measure is some noncounterfactual analog. Comparing against some theoretical external interpretation is counterfactual. Of course, there are some interactions where one is explicitly attempting to better understand a previous message. One could also consider measuring assertions against what one initially thought they knew about something (e.g., learning that the state of a friend's family now includes a newborn). While these would be useful measures, they would not cover a universal magnitude of grounding. It is empirically obvious that grounding can occur simply by activating vectors without any obvious change to one's beliefs about the world. We need a grounding measure that is not based on some change in possible interpretations, which requires an external model (whether about the world or about our own beliefs).

My tentative claim is that amount of Grounding seems proportional to both **Complexity** and the specificity for each quality (i.e., **Quality-Level Specificity**) evoked by messages. I propose that this is captured as proportional to ranges and number of quality dimensions associated with a message (more accurately, the grounding of a single symbol within a message).⁴⁰¹ A larger grounding evoked by a message should involve a larger number of quality dimensions⁴⁰² but on average, the range of possible values along each quality dimension should be smaller. Assuming this can be captured in a single scalar value, I will call this '**Semantic Specificity**'⁴⁰³ for what follows.^{404,405,406}

physics and other knowledge areas (perhaps even folk psychology). Determine how the claims for different force dynamics processes for different conception (e.g., prepositions, verbs, etc.) can be related to interactions with the real world or simulations.

⁴⁰¹ I will elaborate on the portion of a message that grounding is measured against later.

⁴⁰² This could relate to the Minimum Description Length (Minimum_description_length n.d.), however this length would just be the same as the length of the number of quality dimensions unless any of them have redundancies or have no Quality-Level Specificity (i.e., they play no role in the grounding).

⁴⁰³ It could just as well be called 'Semantic Complexity'. It would be better to have a different term altogether for this measure to avoid conflating it with the Complexity and Quality-Level Specificity values used to measure it.

⁴⁰⁴ These measures assume that a given grounding is a thing. Grounding could be better understood as a process. However, this process cannot simply be the delta between the previous and current grounding experienced (e.g., Semantic Specificity as time t minus Semantic Specificity at $t - 1$). This assumes too much about how a past grounding relates to a current one. Instead, it requires understanding the reception of messages to itself be a process, such that how they are grounded is also a process. This could require a considerable overhaul on the theoretical apparatus applied for understanding Semantics and Information. I have no clear reason at this stage to make such a shift, so claims will be described as if assuming messages and grounding are things.

⁴⁰⁵ *Research: Even if it is not the appropriate Grounding measure, consider what semantic property might be measured as a change from one state to the next. For example, a change in grounding could be measured by differencing vectors. What role should convexity across vector sequences play? A subtraction over a shared convex region leaves only what vector properties have some shared meaning with the previous vector, but differ, such as an assertion of a previously unknown property about what the vector represents. Also consider how messages might first increase the number of possible interpretations and then reduce interpretations. The trajectory from one hypervolume to another would then be what is experienced, which would differ depending on location and shape in Conceptual Space. Such a measure is dynamic, and thus active in a sense that could distinguish what is experienced (rather than the other unchanging areas).*

⁴⁰⁶ *Research: The concepts discussed in this chapter still need to be formalized and tested. Determine how to measure Semantic Specificity for existing AI models and how it could be leveraged for them.*

Greater Quality-Level Specificity as a less uniform distribution over each given quality dimension's value range would generally imply a less Shannon entropic macrostate, while more variables represent more possible outcomes, which would tend to increase Shannon Entropy. Complexity is associated with the number of Quality dimensions needed to describe a given vector, while Quality-Level Specificity is associated with average Shannon Entropy⁴⁰⁷ across those vectors. For Semantic Specificity, we can take a weighted sum of the quality dimensions, where each quality dimension is inversely weighted by its Shannon Entropy.⁴⁰⁸ The rationale is that greater Grounding could be associated with more qualities involved, but smaller value ranges for a quality are more pronounced and meaningful (i.e., more poignantly experienced). A final requirement is that only activated vectors within a convex subspace are considered in a single Semantic Specificity measure. The grounded space must be convex for the reasons described in (Gärdenfors 2014). Putting these together, Semantic Specificity as a measure for Grounding / Semantic Processes is: **Given a message, Grounding is proportional to the number of qualities (dimensions) activated within a convex subspace, with each quality weighted as inversely proportional to the entropy over the Shannon Entropy of its active values. If multiple convex subspaces are activated, then they must be measured separately, as they would be grounding different symbols (even if simultaneous). The span of the set of activated vectors defines each convex subspace.** This should capture Grounding magnitude as the Quality-Level Specificity and Complexity of Interpretations of message elements.^{409,410,411}

Number of dimensions should relate to discernment, in that, certain dimensions may be needed to separate possible classes/categories for messages (e.g., size as a quality dimensions for a toy Ford Explorer vs. a real Ford Explorer). Also, Quality-Level Specificity of a given quality dimension relates to how discernable a difference in its values is for a message element. If large changes in a vector for a given quality dimension cannot discern different messages (or perhaps vice versa), then the grounding is less specific.⁴¹² Intuitively, if one receives a word they have

⁴⁰⁷ Technically, the measure must be able to handle continuous distributions for some quality dimensions.

⁴⁰⁸ It might be best to scale this such that the reciprocal of the maximum entropy for a given dimension equals zero. Intuitively, a uniform distribution across a full quality dimension range adds no grounding to the vector. One might also want its maximum to be set to one, but this requires further consideration.

⁴⁰⁹ Note that this proposed measure should only be considered as something to start the conversation. I would not be surprised if it turned out to be deficient. If so, I hope this leads to more appropriate alternatives.

⁴¹⁰ *Research: An alternative but counterfactual measure to consider is to compare the actual entropy over all dimensions vs. the maximum possible entropy over all dimensions. This way, number of quality dimensions activated still matters (because there are more components to add to the sum), as does their Shannon Entropy (since this determines how much each dimension variable contributes to the sum). Although this is counterfactual, it could be useful as a surrogate measure for something like Semantic Specificity.*

⁴¹¹ One might wonder if more dimensions, which increases the hypervolume, should actually be less grounded because it would seem to increase the number of possible interpretations. However, the number of qualities should really be thought of as more specific interpretations and measured accordingly.

⁴¹² *Research: Consider how this could be measured in humans (i.e., self-reports) and AI (e.g., changes in predictions). Also, if the Semantic Vector Space comes from generative model, then less overall Semantic Specificity would tend to generate less-specific outputs. Note that, whatever the type of model, if evaluating Semantic Specificity based on outputs, then how outputs are generated must not misrepresent the amount of grounding happening internally (e.g., the grounding of the neural layer for the Semantic Vector Space being assessed vs. downstream outputs observed).*

never heard before, then they have no way to determine a specific place where it belongs (except at the level of the phonemes received). Additionally, if an area around a vector is undifferentiated (i.e., too few past observations used for learning, such that the subspace is not dense enough to evoke rich grounding), then moving the vector around in it will give the same interpretation for potentially different messages. This also means that the same message could be placed in more possible conceptual space regions, and therefore a larger interpretation space. Therefore, a less grounded space results in a lack of specificity for messages received and more messages that will have the same grounding. Whether Shannon Entropy is the appropriate measure to capture this must be determined. Specificity could also be associated with the density of the quality dimension. Density as a factor for Specificity of interpretations is consistent with the increased knowledge within a Semantic Vector Space model that has greater coverage (i.e., span and density) from training data for a diverse array of observations to evoke richer Semantic Processes.^{413,414,415,416,417} Finally, if there is limited fidelity in neural network weights, then a vector within some distance may be experienced as the same. Thus, lower fidelity should generally be associated with less Semantic Specificity as well.

Let us consider if the qualitative implications of this measure of Semantic Specificity for Grounding magnitude aligns with intuition.⁴¹⁸ First, consider a receiver that only knows of two animal types, dogs and cats. Any messages received, including ones of other animals, will be interpreted only with cat- or dog-like properties.⁴¹⁹ This should have low Complexity, because many dimensions will be unavailable (e.g., flight characteristics). Quality-Level Specificity, while lower in some cases due to lack of observed variation, could be somewhat large within certain ranges for some qualities (e.g., playfulness). Now consider a case where a receiver has observed many animal types but few examples of each. In this case, Complexity could be high since many

⁴¹³ **Research: Determine what density really means for each quality dimension produced by a neural network.**

⁴¹⁴ *Research: Consider if the grounding of activations should somehow be measured by the number observations that have been associated with the surrounding space of an activated vector, each weighted by distance, with closer observations being more impactful. Each quality dimension could be weighted inversely proportional to the length between the two closest concepts with positions on that quality dimension that are assumed to be different from the concept currently being assessed with this proposed measure. More generally, this length for each quality dimension may not only be proportional to the two nearest concepts that are different than the current one, but some more complicated measure based on the density of unlike points for concepts for each given quality dimension. Is there a way to do this that would not seem arbitrary?*

⁴¹⁵ *Research: Determine how to confirm that this grounding measure is consistent with a receiver's behaviors. Model an agent with a frequently updated semantic vector space as new information is received and used in some way.*

⁴¹⁶ *Research: Consider what tests could be performed to compare ERMs vs. Semantic Vector Spaces given this measure. Consider possible controls over a variety of performance tests, such as iteratively reducing their Semantic and Effective Processes (described later in this chapter) to measure the effect on results. Consider if there is a way to also compare to human results. Semantic and Effective Processes could not be measured directly in humans. Could we capture anything related using fMRI?*

⁴¹⁷ *Research: Consider measuring grounding for a symbol within a certain time window to capture how grounding unfolds.*

⁴¹⁸ The examples that follow are greatly oversimplified. They are only meant to provide a basic sense of what is being proposed.

⁴¹⁹ I would think that one needs at least two possible values for a quality dimension for that dimension to have any quality.

different qualities might be available and activated. However, Quality-Level Specificity would be low, assuming many observations are required to reduce Shannon Entropy for the relevant quality dimensions (i.e., those that would end up being less uniform with more observations of the various animal types).^{420,421} The proposed measure assumes that, in both of these cases, magnitude of grounding would be considered relatively small compared to our own and interpretations are generally vague and/or broad.⁴²²

A more grounded convex space generally should have more dimensions with denser qualities represented. This should in theory represent the ability for more specific interpretations of messages for micro and macrostates. Generally, composition of symbols should have higher grounding since they are both more specific and complex (e.g., a gray dog should be more specific than gray or dog alone).⁴²³ Also, larger subspaces might represent more abstract concepts (e.g., Blue is more abstract than Sky Blue and covers a larger space within the color domain). These subspaces for concepts must remain convex to be a single grounding/interpretation. Note that grounding in this way assumes a reduced Specificity for more abstract concepts. However, it is often the case for humans that more specific prototype subspaces are activated with abstractions, perhaps with some qualities only relevant at the given level of abstraction (the latter associated with implicitly knowing that the symbol grounded is meant to be an abstraction rather than the prototype instance). Regardless, this should not be taken to mean that abstraction is less *effective* in some cases, which is a separate measure and discussed in the

⁴²⁰ Note that a comparison across theoretical recipient models does not contradict Semantic Specificity itself being non-counterfactual. What I want to avoid is the process of semantics *for a recipient* requiring counterfactuals, not a third-person measure designed for comparing across receivers (or the same receiver over time). Semantic Specificity is an external measure. It is also an external measure about the grounding itself, unlike Semantic Information Theories (discussed later) which measure over the assumed symbolic products of grounding.

⁴²¹ Why only a few observations of different values for a quality is less specific requires some explanation. The assumption I am making is that variety is required to contrast values. If a (*tabula rasa*) model has only experienced yellow and red, then the color spectrum for them consists of *A* and *not_A*. Until new colors are experienced (and grounded), any messages could only be grounded with this binary spectrum.

⁴²² *Research: If Semantic Processes measure degree of grounding, can we claim that they measure (along with Effective Processes to constrain them) to what degree observations bridge a Description Gap with respect to an observer's model? If multiple reference frames / levels meet the grounding criteria, then can we use the outputs of these layers to compare them to be able to measure something about the associated Description Gap?*

⁴²³ This is also consistent with the Semantic Information Theories discussed later.

sections following the Semantic/Grounding and Semantic Information sections below.^{424,425,426,427,428,429,430}

Convexity allows for many operations to be valid with respect to a continuous change in grounding given distances in the Conceptual Space (Gärdenfors 2014). Unlike convex subspaces, non-convex subspaces cannot result in unified interpretations, and so they also cannot be considered a single abstraction. From personal experience, when two or more exclusive interpretations are possible, there is often an experience of ambiguity or lack of comprehension. This suggests that the grounding not only directly relates to Quality-Level Specificity but that

⁴²⁴ *Research: Changing scales often changes symbols and what they represent. When humans do not predetermine the mapping, Semantics is needed to select macrostates for the emergent properties across scales. Effective Processes should constrain propensities for what properties to smooth out or change altogether. In what cases, if any, is the optimal abstraction level (i.e., macrostate/microstate) for the message and the Knowledge System the one that maximizes Grounding? What about minimal Grounding?*

⁴²⁵ *Research: If less grounded concepts have larger ranges along quality dimensions, then how would this be differentiated from abstraction of concepts? Is a mammal always less grounded than a cat? Or are there particular dimensions associated with abstraction? Or perhaps a larger quality range is only appropriate for abstraction, and less certainty for an object is associated with some measure of salience, confidence, or weight associated across vector elements? Are standard ANNs too artificial to appropriately distinguish them (i.e., certain properties of biological neural networks are needed)? Is a percept that cannot distinguish a sensed dog or cat the same thing as a dog/cat abstraction? It does not seem to be. Instead, there seems to be a quality of uncertainty. If so, then this may require modifying the Semantic Specificity measure.*

⁴²⁶ *Research: Look for existing research that explains what performance varies with less training observations for a given subspace. What does this effect have to do with the size of an interpretation space for which semantic information is supposed to measure reduction? How does subsequently learning new facts improve future interpretation ability in humans? Is it adding new information in something like a ‘fact of the world’ dimension that provides the additional specificity?*

⁴²⁷ *Research: In many Deep Learning architectures, such as Convolutional Neural Networks, deeper layers learn to represent more abstract features. What happens to semantics as one compares the semantic vector spaces across these layers?*

⁴²⁸ *Research: Even though Blue is more abstract than Sky Blue, it does evoke a specific blue. That is, perceptions often seem to evoke more sensory details than the messages themselves constrain (e.g., the message “dog” evoking the visual perception of a specific dog). How does this relate to interpretation and why does it happen? Is this specificity unique to sense-based quality dimensions, such as vision, because vision cannot be tuned to receive anything but some certain level of specificity? Is there no way in principle to perceive anything at a higher level of abstraction? Is the issue that abstractness is only a property of a concept? Consider if the answer is that the construction of a sense evoked by the Conceptual Space always involves selecting a point in a region, but selection is accompanied by the understanding of the concept being at the level of the region associated with the abstraction level tied to the symbol(s) received. That is, there is a dimension for a quality associated with the appropriate abstraction level.*

⁴²⁹ *Research: Consider how a new message can be semantically perceived as already known, perhaps because it is below some perceived resolution of difference for a known vector, and yet be experienced as novel. It is interesting that we can both generalize new situations while maintaining implicit understanding that the situation has novel elements (such as simply being a new episodic event). Do we never truly experience the same grounding more than once due to subtle changes in context (e.g., lighting in a room, the temperature, or what the agent was just thinking about)? Do we have some vector distance threshold to decide when stimuli should be perceived as the same (e.g., seeing a friend after a year and still experienced as the same person)? A threshold may be associated with an attractor guiding certain inputs closer to certain established vectors (i.e., close enough to be below the semantic fidelity). This aligns with the relation of Conceptual Spaces to Prototype Theory (Gärdenfors 2014).*

⁴³⁰ *Research: Consider if there is a way to measure the amount of semantics/grounding in models by controlling for syntax (i.e., replacing terms without changing syntax). What would be measured? What would the results mean?*

there is also potentially a quality dimension or domain for grounding the experience of Ambiguity itself, which might be activated when competing interpretations remain present.⁴³¹

Semantic Specificity could be involved in resolving tensions between activations. Exclusive activations naturally inhibit each other, such that they are not fully “activated” until one is chosen. This is consistent with Gestalt phenomena, such as figure-ground oscillations (e.g., Rubin’s Vase (Rubin_vase n.d.), Necker Cube (Necker_cube n.d.)) and logical paradoxes.^{432,433,434} More generally, Semantic Specificity could be involved in the search process applied to determine stable grounding for message elements. As discussed earlier, an underlying ANN should have attractor states. Each attractor state could represent a semantic distinction (i.e., unique grounding) for symbols received within messages. There could be very sophisticated processes involved that could include intrinsic use of Semantic Specificity values in the search process to reach the attractor with the “desired” grounding (which includes the “desired” emergence level).⁴³⁵

We must also determine what exactly for a message should be measured for grounding. It seems that it should be some instance of the message, and thus related to a single symbol. However, context, including previous symbols within or across messages should influence that grounding, including increased Semantic Specificity of the meaning of a message as it is received. However, a full context switch (e.g., starting a new conversation) would be an odd comparison. Thus, what is symbolized is not necessarily aligned to a single message element.⁴³⁶ A full sentence could be the symbol. This makes it difficult to measure but is more natural.⁴³⁷ One can choose to

⁴³¹ *Research: It also suggests a need to consider how multiple separate convex subspaces activated at the same time might reduce overall grounding. Consider if a salience measure is needed to weight the total Semantic Specificity, which, like a probability distribution, must be distributed across activated non-convex subspaces.*

⁴³² *Research: Determine what an appropriate measure for Stability would be to associate with Grounding.*

⁴³³ *Research: If nonconvexity creates this tension, including for bistable phenomena, then consider how convexity relates to neurophysiological concepts such as attention shifts and neural “fatigue” (Pitts, et al. 2009).*

⁴³⁴ Note that the fact that we still know that these are paradoxes suggests we map the oscillatory behaviors of these examples to a higher level that grounds these as paradoxes. This suggests that we have learned or have innate attractors as higher-level patterns to map such types of dynamics to.

⁴³⁵ *Research: How do new symbols become grounded? Consider Peirce’s Semiotic theories, including for the nonsymbolic signs (Deledalle 2000). Iconic and indexical signs become associated with referents based on similarity or causal relation, respectively. Conventions, including being told how to map new symbols to concepts to models can be provided by others by using already grounded symbols to explain how to ground new symbols. This would be an important area of study for exactly what must be communicated for a receiver to ground a new symbol so it can be applied to computer models without having to provide them an initial dictionary of symbols. Indeed, this question arises for any symbols. Also, consider if this is one of the roles for phenomenal experience. If a receiver does not have the sense associated with comprehension for a message, then they have not located a model to apply it against. This could be because they have never received the symbol(s) before and embedded them in conceptual space or because they have forgotten the process that must take place in the vector space to get to the appropriate vector location.*

⁴³⁶ There is no reason to assume a symbol must explicit map cleanly to parts of messages. A symbol at a given moment might even overlap across parts of the same message elements.

⁴³⁷ *Research: Consider how grounding relates to the number of quality dimensions that must be reconfigured given a message. Additionally, the larger a sequence that can be grounded in a single but more configured vector space, the richer and more semantically compressed is the process given that message. This might be maximized and is limited by a point in which no more discernable dimension changes can be made, thus losing the ability to distinguish messages beyond that point. This may, for some sequences, require that the message element being grounded be*

compare groundings for different models against a predetermined set of different message partitions, compare to partitions humans believe are single groundings, determine how to take a maximum grounding, or a variety of other options. However, the more natural option is to allow Effectiveness to constrain how the grounding propensities form for a given agent. Temporal dependence might also need to be considered, not just sequences.

Such a grounding measure can also relate to the micro and macrostate since it determines the indices. **I tentatively argue that the vector subspace activated is equivalent to the micro and macrostates. If so, then the micro and macrostates are more of a theoretical construct and are always implicit (except in the case where the implicit properties form explicit micro and macrostates). This means that the magnitude of Grounding is equivalent to the Semantic Specificity of the micro and macrostates. Since any explicit micro and macrostates are only part a more foundational implicit grounding, they are not necessarily correlated with the Semantic Specificity of the given grounding.**⁴³⁸

Assuming that Semantic Vector / Conceptual Spaces are appropriate models for measuring Grounding, what does it mean to “activate” a vector or subspace given a message? In a Semantic Vector Space generated by an underlying neural network layer, the activation is simply the vector produced by the activated nodes of that layer given an input message. How value ranges would be associated with activations in the neural layer(s) associated with the Semantic Vector Space requires further consideration. Perhaps a requirement is that activated values for a given quality dimension must be associated with a distribution over neurons/nodes.⁴³⁹ For example, if a message received is ungrounded, such as hearing an unfamiliar or ambiguous word, such that the receiver does not comprehend (understand) it, then the receiver’s semantic vector space is “uncertain” for where to place them. In such a case, there might be a more uniform spread of activations across neurons. This could happen when the vector space has received a sparse set of observations or when whatever the process is for finding the vectors given messages is faulty. It will also often depend on the semantic clarity of the message itself. This could require a different approach than assuming each neural node is a Semantic Vector Space dimension. It is

split up for separate grounding. However, this must not be thought of as necessarily relative to one’s previous state. Instead, it could be the complexity of the macrostate formed, but in such a way that larger ranges of possible values for each given quality dimension used to form the macrostate lowers the magnitude. This makes sense because the more specific a value is on a quality dimension, the more reduced it is.

⁴³⁸ Note that measuring Shannon Information still requires an explicit micro and macrostate to measure how emerged symbols associated with the grounding update them. How this might be associated with the grounding (e.g., change in hypervolume) must be determined, but this is consistent with the notion that grounding is required to actualize Shannon Information (e.g., a volume is required to measure volume change).

⁴³⁹ Bayesian NNs (Goan and Fookes 2020) measure uncertainty with predictions.

also more intuitive, since it is implausible that the value for single neuron would be responsible for determining, say, the hue of a given color received.^{440,441,442}

Another phenomenon that should relate to Semantic Processes is **Coherence**. The models we form might be driven by dynamics towards attractors that “seek” interpretations of message that are stable in some way with our existing understanding of our world.⁴⁴³ This is consistent with some, common but difficult to describe, experiences associated with how we ground messages. Thus, Coherence may be a separate measure than Semantic Specificity, but there is some implicit

⁴⁴⁰ *Research: In theory, another way to measure Semantic Specificity would be the number of messages that can be distinguished by the system. This would also presumably relate to something like “Semantic Capacity”. However, it must be determined what this would mean for in continuous metric space. This could be estimated in some way by the similarity of model activation dynamics for different messages. If two semantically different concepts evoke the same model dynamics, then this reflects lower semantic capacity. Since this requires comparing grounding for messages, it is counterfactual, and thus external, but perhaps useful as a standalone measure. I would expect it to correlate strongly with the proposed measure of Semantic Specificity. Also consider if the magnitude of semantics/grounding might be proportional to how distinguishable the system (e.g., neural) state is for messages. As a counterfactual measure, if two messages result in two different neural states, then one can say the messages were distinguished. This captures the impact of system fidelity on grounding. A related measure is differentiation of the model. If two regions have all the same neural properties despite being associated with different quality dimensions, then there should be no difference in their grounding properties, even if they are activated by different messages. While identical regions would be highly unlikely, it points to the possibility of considering how similar different regions are in determining grounding properties.*

⁴⁴¹ *Research: Different NN layers can be used to represent a Semantic Vector Space, and thus determine grounding potential. For agents with multi-layer ANNs, it is unclear what layer should be associated with grounding. Are there separate Semantic Processes occurring or must they be considered a single one? Something about the question here itself seems artificial. Criteria similar to those used in Integrated Information Theory (i.e., $\max \Phi$) (Tononi, Boly, et al. 2016) could be considered to make the determination (I discuss IIT later in this chapter). That is, how $\max \Phi$ is used in IIT to determine what part of a system is conscious might also be applied to determine grounding regions. However, I am skeptical that it makes sense to choose one region as grounded over others, rather than them all being separate groundings. It seems more natural to instead force that any separate groundings are grounding different symbols (as discussed in another footnote). Even this only makes sense when there is actually separation across layers. Given that layers will typically exhibit dependencies, the most natural interpretation is that they form a single vector space, where convexity would be used to determine separate groundings. More generally, consider how IIT could be applied in such a way that it can be made consistent as a measure within Conceptual / Semantic Vector Spaces.*

⁴⁴² *Research: Determine how Semantic Specificity can be measured for different vector space models and if/how it can be useful as an optimization function for them.*

⁴⁴³ I have in mind here the types of attractor dynamics found in models, such as how Hopfield Networks (Hopfield 1982) complete incomplete or noisy messages.

relation between them (e.g., Coherence should constrain interpretations). It could perhaps include some cognitive error correcting process.^{444,445,446,447,448,449}

Maximizing Semantic Specificity would often result in over-specifying/overinterpreting (or complexifying) messages. Effective Processes (discussed later) constrain the particular Semantic Processes that form, which does not necessarily involve maximizing semantics. Gestalt Psychology (Todorovic 2008) suggests that minds often construct the most stable, and often simple, objects from scenes. What is true for visual perception is likely true in kind for cognition overall. This tendency towards simplicity is associated with abstraction, which helps cope with certain types of messages by grounding them as some other semantic whole. It is often more effective to behave the same way for similar environmental objects or events and to interact with who systems rather than perceiving each individual component. Gestalt examples, of course, may be best viewed as principles constrained by evolution, such as making out objects that are partially occluded in complex ways. The challenge is to determine if there are any general measures for Grounding that can be constrained by evolution to create derived principles, such as totality, stability, etc. (Todorovic 2008).⁴⁵⁰ Effective Processes should characterize such constraints. From an evolutionary perspective, there is some optimal level of specificity for different scenarios (i.e., how much Semantic Specificity helps cope with different

⁴⁴⁴ **Research: Determine how Coherence would be measured and how it relates to Semantic Specificity.**

⁴⁴⁵ *Research: How does Coherence relate to Comprehension? How can this be modeled in AI? Is there some way we match new data semantically to the existing semantic Knowledge System, not only by semantic distance but by some structural reasoning process or processes? Once we can build such a semantic model then we should explore Cognitive Science-based methods to reason with it.*

⁴⁴⁶ *Research: How are new prototypes created in Conceptual Spaces? What causes the mind to decide when new stimuli warrant a new domain or prototype (not just shifting a prototype)?*

⁴⁴⁷ A related concept is Harmony, which is formalized into a Harmonic Grammar in (Smolensky 2012).

⁴⁴⁸ *Research: If Semantics captures knowledge that can be used across many macrostates, then the amount of semantics might need to capture this cross-model knowledge gain. This relates to transfer learning, as the subsymbolic grounding of messages has theoretically unlimited potential to increase knowledge given future messages. How to capture this must be determined.*

⁴⁴⁹ **Research: Nothing in this section explains how a specific quality dimension obtains its unique quality. If one dimension was swapped for another but the overall vector space geometry was preserved, why would the grounding change? Are there some innate/primary qualities? If so, then how do they obtain their character? How are the dimensions themselves grounded/learned? Can new dimensions be created, and if so, how? Consider the possibility that concepts experienced more frequently and in different varieties can end up (conceptually speaking) moving from a position in the conceptual space to being a dimension itself (or more speculatively, some holographic representation, such that the vectors somehow simultaneously determine the dimensions they are embedded within). For example, when one sees a cat for the first time, they would understand it as some sort of mix of other known animals (e.g., interact more like dogs, but small and sleek like otters). At some point, though, one can imagine a quality space for feline-ness. What neural topology would support this?**

⁴⁵⁰ **Research: Determine (theoretically and/or empirically) how known Gestalt properties can be accounted for with Conceptual / Semantic Vector Spaces, and what the resulting implications are for the Grounding claims made in this section.**

circumstances).^{451,452} Effectiveness influences not only how, but how much Grounding increases or decreases to ground different messages. Optimizing Grounding could also relate it to Semantic Information Theories, such as defining a separate measure that can depend on other factors, such as Truth (Floridi 2019). Through Effective Processes, maximizing Semantic Specificity can also be traded off by accounting for the reduction in message length that the level of specificity offers, which relates to the function of Semantic Compression discussed later, including reduced energy costs. For example, Specificity would generally be optimal if it captures the full complexity of what a transmitter intended to evoke with the message sent.⁴⁵³

Grounding as Understanding/Comprehension

Grounding as Interpretation of messages can also be described as an agent's Understanding, how they Comprehended the message, or associated with the Meaning evoked by the message.^{454,455} It may differ from some external objective or consensus, and it may change over time.⁴⁵⁶ These external factors are important but provide no insight into the mechanics of how meanings fundamentally arise and can be modeled. While any grounding could be said to be a comprehension, understanding, or meaning, often, for humans at least, there is an active process by a receiver to achieve a certain understanding/comprehension that they feel is correct. Such recognition is part of the grounding properties. Thus, unknown messages might be grounded at

⁴⁵¹ To be clear, the quality of groundings has a larger impact than simply the quantity associated with Semantic Specificity. Effectiveness should constrain both.

⁴⁵² *Research: Determine what beyond simply "Effectiveness" can constrain Semantics (i.e., model forming). A host of interesting regularities might be found with empirical study. For example, we can relate Model Forming to how humans select categorization levels for messages. There should be something about the level of dog for why it is chosen over mammal or big dog, dog with pointy ears, Poodle, etc. for a given image. (Rosch and Lloyd 1978) referred to the cognitive choice by defining the chosen (basic) level as the one that maximizes cue validity; dog has more informational content (i.e., substance) and is easier to categorize. But what determines cue validity? Frequency of observation likely plays a role (i.e., given a picture of a dog, it is the dog level that tends to matter for a given situation), but how does one then quantify what has "mattered" in the past and/or what intrinsic properties "matter" for a given problem? For supervised machine learning the answer is that we explicitly train it to do so. For unsupervised machine learning, there must either be a predefined number of categories or the number of categories changes to optimize some predefined function. Again, while this points back to Effective Information, there is room to flesh it out more with studies, and it should help to constrain how we train AI to emerge similar Semantic Processes as humans for the same messages. There is also considerable potential for future studies to dig into whether there exist some universal "laws" or at least tendencies for optimizing Semantic Processes and Information. For example, there is likely some relation to forming micro and macrostates with maximum entropy to increase expected information gain, at least when a receiver is exploring their world and the actions they take in it. This would seem to correlate with free energy minimization (Friston 2010).*

⁴⁵³ *Research: Consider how a Semantic Processes measure might quantify how current embedding models capture less than humans (at least, from a human's perspective), and yet capture more grounding than an ERM equivalent. Measuring this difference would seem to be a useful metric for Grounding.*

⁴⁵⁴ Some of these concepts may be more appropriate to distinguish from each other, such as Meaning being what Grounding/Understanding achieves. Comprehension is also typically used to imply a message was "correctly" understood. However, the possible differences I came up with seemed open to interpretation and would distract from the main points of the section.

⁴⁵⁵ I will use these interchangeably below (wherever I think one is best for reader clarity).

⁴⁵⁶ Indeed, a strong cognitive-sociological constraint is the iterative process of using social feedback to find common grounding for symbols.

the phoneme level, but the receiver recognizes that the transmitter likely meant for the message to refer to something about the world and know they lack comprehension at that semantic level.

A message that evokes an already existing form can be said to have been comprehended. This implies that, unlike for Shannon Information, Semantic Process / Grounding is nonzero for repeated messages whenever a comprehension event occurs for the receiver.^{457,458,459} A fact about the world that is already known, including declarative facts (e.g., a dog is a canine), provides no Semantic Information according to the theories described in a later section, but it does evoke Semantic Processes. Grounding/Semantics are still experienced with repeated assertions.⁴⁶⁰

Understanding a given message could directly activate existing vectors (e.g., “soup” or “hello”) or could involve inference operations (“spoon is to soup as X is to pasta?”). The former is typically not a fact asserted about the world, while the latter either outputs a known fact or reasons to a new fact that was in the Conceptual Space as a potential fact now actuated.

⁴⁵⁷ I am referring here to messages repeated for the same outcome, not two outcomes that are the same but give unique evidence for some belief. The latter would also be nonzero for Shannon Information.

⁴⁵⁸ *Research: Interestingly, repeating something many times in a row can lead to a seeming loss of comprehension, such as when a repeated word starts to seem foreign. In such a case, I believe it is appropriate to say that the word (i.e., symbol) loses grounding. Presumably, this relates to habituation of specific neurons or neural circuits. If so, we should be able to measure what neural processes are associated with, and perhaps causal to, semantic processes (i.e., grounding). This would be an interesting study, perhaps with insights for how we model Conceptual or Semantic Vector Spaces for AI. The process of attempting to understand something is also worth investigating. If it could be identified in humans, it could potentially be incorporated in AI as a search process in Semantic Vector Spaces. In the case of reading something repeatedly until it loses meaning, it seems that we try to recapture the meaning by changing our thought process. That process feels like actively searching for an alternative way to ground the parts of the sentence, sometimes in parts that are put back together (maybe at a higher neural layer) and maybe sometimes as a whole (usually by searching for a single gist). This process seems no different than the process used for complicated sentences, which is more common and therefore may be why we evolved this process as adaptive. This process is one of trying to achieve comprehension. The inability to stick previous or current parts of the sentence in memory long enough to bind them with subsequent parts is related to the issue of trying to ground a sentence repeated many times and losing comprehension. This would affect an ANN trying to use it in conjunction with other parts for higher level processing, and it would also impair understanding subsequent parts of the sentence at the same level/layer by not providing context. Note here that context is critically related to grounding too. Experimenting with this process might help understand Comprehension more deeply. What is the brain doing during this process? Reduction in semantics may also be associated with the way many processes when mastered become automatic. Do processes that become automatic retain anything we would want to describe as “semantic”?*

⁴⁵⁹ This claim for how Understanding is evoked differs from symbolic alternatives. For example, (Chaitin 2006) claims that understanding is a form of compression, which is associated with modeling phenomena with a more compact set of rules. I believe this captures an important aspect of Understanding. Semantic Compression is a key part of Semantics and therefore Understanding, as discussed in a section below. However, like the Semantic Information Theories described in a section below, it leaves out how understanding is evoked by the rules. Again, this is difficult to explain, since we communicate in symbols, and there is therefore no way to describe or define Semantics except in an indirect way that converts it into Shannon-like Information. I therefore can only hope to evoke the subsymbolic essence of semantics in the reader through the indirect use of symbols.

⁴⁶⁰ Again, Redundancy here does not include receiving the same measure for different outcomes, which are informative as samples for updating evidence for some associated phenomena.

A lack of understanding of a message is experienced when following instructions or repeating something but not understanding them.⁴⁶¹ This is what it means to parrot something.⁴⁶² When some satisfactory/stable understanding⁴⁶³ is achieved (i.e., the experience of an “aha moment”), the grounding is experienced as if it has been appropriately aligned within one's existing understanding.

Understanding/Comprehension may be its own type of sensed modality.⁴⁶⁴ There is something very fundamental about it. There is a clear difference in any perception going from not understood to understood (or more rarely in the opposite direction, as experienced when repeating the same word to oneself over and over).

Note that **Understanding enables but does not necessitate the ability to explain what is understood explicitly.** I understand how to ride a bike but cannot sufficiently explain it to teach others everything I understand about it so that they could ride a bike as well as me. Grounding/Semantics should not be considered limited to associations with natural language symbols. All senses can be grounded and therefore have semantics. Some processes are likely grounded innately, as well as learned prior to language development, such as the grounding of Naïve Physics (Hayes 1979) and Naïve Psychology (Ruble, et al. 1988). This is consistent with the claim that Semantics is subsymbolic, as it is unclear how to explicitly represent many of these grounding phenomena.⁴⁶⁵ While ERMs would therefore be challenged to model such semantic

⁴⁶¹ More accurately, it is not grounded at the level of the world-meaning of the words. There can be grounding at level of letters or phonemes, as well as a grounding even associated with not understanding the message. These forms of grounding are likely to have less Semantic Specificity and/or an undesired grounding state for us.

⁴⁶² *Research: Given that messages can be grounded as words, phonemes, light and shade patterns, etc., what makes the brain seek to further ground them at a “higher-level” semantic understanding? Consider if part of the method for this involves finding a tradeoff between maximizing Semantic Compression (discussed later, including Mental Chunking) of messages and maximizing Semantic Specificity. This is not the full answer but may be interesting to try to optimize for AI models for which using some Effectiveness maximization measure is infeasible.*

⁴⁶³ I do not mean to suggest that there is always some explicit evaluation of how messages are grounded. By “satisfactory”, I simply mean that there is an implicit experience of “getting it”. By “stable”, I mean something related to implicit stopping criteria or a basin for the search once a certain grounding is achieved. Both require better formulations.

⁴⁶⁴ *Research: How this claim squares with grounding itself being associated with how sensory events are perceived must be considered further.*

⁴⁶⁵ This is a good point to clarify how my claims should be understood with respect to debates on current Deep Learning vs. the need to account for innate cognitive properties. I fully agree with claims that current Deep Learning models are missing almost all innate capacities required for human-like intelligence. What I have emphasized is how evolution adapted such constraints through interactions with the world, and that similar effectiveness-based constraints are needed for AI models. Current training approaches are more similar to controlled teaching episodes (i.e., neither evolution nor lifetime learning). Where I diverge with some is in how such innate capabilities/constraints would be implemented. One might reasonably challenge the need to go through evolutionary processes when many of these innate constraints are known. I would agree that we should determine how we might effectively implement such constraints rather than have our models attempt to rediscover them. However, I am arguing that implementing these innate constraints must be done at a subsymbolic level. Directly programming in symbolic constraints would divorce them from everything else that must be grounded through subsymbolic processes. Given their importance, these innate constraints should not only be grounded to operate effectively but should affect the grounding of most phenomena, including what is learned and how. Explicit approaches would also only account for a small number of the actual innate constraints

processing, a variety of research projects have already demonstrated the use of embedding methods to ground nonlinguistic tasks, such as in computer vision and learning agent policies (Lynch, et al. 2020, Yan, Zhang and He 2019). It is therefore as appropriate to say we understand images, sounds, proprioceptive stimuli, etc. as saying we understand natural language.

Finally, while I have described Semantic Processes / Grounding as Comprehension, Semantic Information Theories tend to measure how informative messages are, which can be described as Learning.^{466,467} How do Semantic Processes relate to Learning (i.e., knowledge capture)? Learning changes future interpretations (in the Comprehension sense) for messages. This could be thought of as “**grounding modification**”, which creates some short- or long-term change in underlying neural properties and therefore to the semantic vector space. In turn, this can result in updated explicit knowledge by modifying the properties of macrostates for future messages.⁴⁶⁸ Unlike Semantic Specificity, Semantic Learning would be essentially zero for messages asserting known facts, which is more like existing Semantic Information Theories. Modifications could include a shift in the geometry of boundaries encapsulating a domain, adding partitions for new domains, or perhaps even new quality dimensions.^{469,470}

Semantic Information

Introduction

This section comes back to the notion of Semantic Information. As a reminder, I believe Semantic Information has more in common with Shannon Information than Semantic Processes / Grounding. I do not claim these information types are exactly the same. However, all information types can be understood as measuring a change in macrostates given messages and microstates. Semantic Processes are subsymbolic processes involved in forming macrostates, microstates, and associated symbols for messages. The following sections elaborate on this difference.

that exist. Many are unknown and many of these will not be describable due to the Description Gap. Conceptualizing cognitive constraints that form from a massively complex system is limited to what can be interpreted through external humans. Moreover, we may find that AI can in some cases adapt its own innate constraints that make up for inevitable environmental, evolutionary, and structural differences from the way we developed our own. These will not be knowable beforehand. In summary, I agree with those who are concerned about the reliance on learning in current Deep learning. How innate heuristics are implemented must leverage Effective, Semantic, and Information Processes. There is a wealth of knowledge in Cognitive Science and other fields that needs to be applied to achieve AGI, but my hypothesis is that how they are applied matters and should be consistent with various claims in this manuscript.

⁴⁶⁶ Learning here includes learning what somebody meant, temporary states of affairs, etc.

⁴⁶⁷ However, (Lundgren 2019) does distinguish information in messages from how informative a message is.

⁴⁶⁸ *Research: Consider how conditional inference can be performed in a Semantic Vector Space (i.e., conditional grounding), including for positions, subspaces, and dimensions. For example, receiving the message “base” before “ball” constrains the grounding for “ball”. However, complete context switches would not (i.e., switching between two independent discussions).*

⁴⁶⁹ *Research: What neural network changes would be responsible for each of these modification types?*

⁴⁷⁰ *Research: Model “nodes” can either be activated by new data or can be changed with new data. For example, a neural network node can get activated by an input, but it can also get ‘activated’ in a sense by backpropagation or some other form of learning. Must Semantics account for both? What difference is there, if any?*

I start by describing some existing Semantic Information Theories and propose why they do not address the core of Semantics/Grounding. They all focus on some externally measured set of possibilities, such as some objective truth “out there”, the transmitter’s intent, or the receiver’s previous state (or some predetermined “beginning state” N steps ago). They are thus counterfactual and external measures, each of which can be useful for different applications. These theories work fine as external measures of Semantic Information but fail to describe semantics internally to the agent/model, which is where the meaning of symbols is found. They only *seem* appropriate for intrinsic semantics because we as semantic observers of the measures for the systems of interest ground the symbols as representations for the world. They model what the Semantic Processes construct rather than the processes themselves. Since they do not address the core of how messages get their meaning, there is no need to assign one as “the” fundamental theory. They are all consistent with a more general concept of Semantic Information for which scope and level of the microstate and macrostate is open to what a modeler wants measured. Semantic Information could be based on truth, intention, or anything else one wishes to measure. It is flexible because it is an external measure of the system, and an external measure can be defined without having to be some fundamental property of the world.

Semantic Information Theories

Approaches to Semantic Information Theory generally share a focus on reducing semantic ambiguity of some domain given messages. The more that a message reduces ambiguity about some domain, the more semantic information that message provided. This can be equated to reducing the number of possible interpretations and therefore also specificity of messages, although the specificity must evoke factual interpretations for some semantic information theories. While this is similar to what Semantic Processes measure, it no more similar than what Shannon Information measures. A more specific message can be one that provides greater amount of information gain about a microstate. If a message was as specific as the microstate, then the Shannon Information gain has maximally reduced ambiguity about the microstate. Thus, although the descriptions of theories that follow may seem similar in intent to my description of Semantic Processes, they are not special in this regard.

Theory 1: Semantic Information as Logical Probability of Assertions

I start with (Bar-Hillel and Carnap 1953), which provides an early attempt at formalizing Semantic Information. Here, semantic information is associated with the logical probability for assertions made by messages. Unlike Shannon Information, which measures the probability associated with the basic⁴⁷¹ symbols for elements in message sequences, (Bar-Hillel and Carnap 1953) measure what elements in a sequence reference about the world. This can have different information values than the Shannon Information of the basic symbols. Using logical probability,

⁴⁷¹ By “basic”, I mean the elements symbolizing what we tend to consider as not referencing anything else, such as numbers or letters of the alphabet. As I said earlier, I think it is a mistake to see these as fundamentally different. As symbols, they do pick out referents for us, even if they seem less semantic or somehow more contained (perhaps these elements might be explicitly approximated to be symbols that reference their form *as certain symbols*, but this still requires implicit grounding that is unavailable in the element forms alone).

it is straightforward to see that a message asserting *A and B* has greater or equal Semantic Information than *A or B* because *A and B* is always equally or less likely to be the case. Likewise, a message describing something as a mammal is less descriptive than describing it as a dog. A dog is more specific and less logically probable, all else being equal.^{472,473}

Interesting implications have been identified for this theory. One is that the relationship between the probability of claims leads to the “Scandal of Deduction”, where (Hintikka and Suppes 1970) show that any deductive claims have zero semantic information because the conclusion follows from the premises with 100% probability. If deductive assertions are necessarily tautological, then one might expect that they carry no semantic information. However, this is inconsistent with the understanding we often gain through deductive inference. I believe this is resolved by contrasting this external objective measure with how deductive processes impact us as imperfect agents. Deductive inferences are valuable for humans because we do not necessarily know the conclusions until we follow the logical consequences of the given premises. A related issue with this theory (also discussed later) is that a contradiction has maximum Semantic Information, since it has 0% probability of obtaining. As all Information requires an observer, it is not so relevant that an objective measure (that is, measured against some external theoretical world of facts) ends up with zero or maximum semantic information for tautological or contradictory claims, respectively. This warrants also considering a Semantic Information measure modeled relative to a given human’s Knowledge System rather than some objective state of affairs.

(Bao, et al. 2011) provides a Semantic Information model that builds off the approach by (Bar-Hillel and Carnap 1953). They provide a useful example to describe the difference between Semantic Information and Shannon Information. Consider two possible messages:

A. Rex is not a tyrannosaurus

B. Rex is not a dog

(Bao, et al. 2011) associates Semantic Information with the macrostate applied to our world domain (the microstate here being Rex’s animal type), where dogs are not extinct and thus more common than tyrannosauri. As such, *A* is less semantically informative than *B*. Now consider Shannon Information associated with word probabilities. “dog” is more likely than

⁴⁷² Research: Consider how Semantic Information as reducing degrees of freedom within a Knowledge System given messages might be measured as proportional to the deviation in a Receiver’s Knowledge System (i.e., what they believe). For graphical models, this would relate to pruning edges or making their conditional probabilities less uniform for given messages. This might be generalized to the total causal effect size of a message to the network. This links to Shannon Surprisal since less probable messages would be expected to have larger impacts on existing knowledge. As an example, “base” could refer to many things until “ball” is received. Degrees of freedom could be measured with respect to the relationships of each term; including, how many possible relationships are removed with additional data (although it is certainly more complicated than this).

⁴⁷³ Research: Determine how primitive operations (e.g., AND, OR), and how they change degrees of freedom, are associated with changes in grounding properties (e.g., base + ball = a different unary concept than base or ball). How is this semantic change quantifiable? How does Semantic Compression apply to the context of the three different terms?

“tyrannosaurus” because it is more probable given the statistical distribution of English words. *B* is more Shannon informative than *A*. Of course, we could construct a different pair of messages where Semantic Information was high and Shannon Information low.

Why Logical Probability does not cover Grounding: Measuring how much messages reduce logical possibilities can be interpreted as a reduction in the number of possible macrostates describing some state of affairs. However, it does not itself determine how to form the micro and macrostates, and thus first requires a semantically capable external agent to construct them and map messages to them. The theory does not *explain* the role of semantics in determining what symbols refer to; it instead assumes the meaning of symbols is formed for a certain domain. The biggest difference between this measure and the Shannon model it is compared against is simply what the domains are assumed to be (e.g., English words or assertions about the world). Beyond that, only the use of logic-based probabilities differs. They both ultimately measure the uncertainty about symbolic microstates (e.g., a macrostate about *Rex(Words)* or about *Rex(Animal Type)*).⁴⁷⁴ The difference in this uncertainty fully depends on what some semantic *external* modeler decides the symbols reference.

Of course, the *Rex(Animal Type)* microstate seems more semantic than a microstate about how words or letters are distributed. This could be because: 1. The grounding associating elements to the real world is much more prevalent for us than the grounding of elements as words; and 2. The real interpretation / model formation is assumed rather than explained. Neither of these are good reasons to suggest that grounding does not generalize to all model forming processes, even if it is more obvious to us (perhaps for folk psychological / evolutionary reasons) for certain types of micro/macrostates.⁴⁷⁵ More important is that there is nothing in the formulation to account for the grounding of the assertions. This process is necessary to determine the Shannon-like model(s) to apply messages against. A measure that *assumes* messages are meaningful does not address the core of Semantics. Semantic Processes determine

⁴⁷⁴ While it remains to be proved, the variety of possible logical constructions that could result in changes in logical probabilities can be understood as macrostate changes. Consider that the message, “a dog is at the door” is more semantically informative than “a mammal is at the door”, while “mammals are placental”, seems more informative than “dogs are placental”. A Semantic Information Theory approach would account for the differences in their logical structure. However, they can also be accounted for based on microstate differences. One microstate is about what is in a location, while the other is about what is true about some set of animals. Also, note that, had one of every type of mammal been at the door, then transmitting ‘mammal’ would have been more informative. While I will not formalize the “correct” microstates here, the point is that the concept of micro and macrostates seems flexible enough to account for both statistical and logical probabilities.

⁴⁷⁵ (Bao, et al. 2011) describe their formulations as a measure over possible “models”. This seems to mirror the model forming requirement for Semantic Processes, but it is still a Shannon-like Information measure. The key difference is in whether symbols are assumed (Shannon-like Information) or if the receiver itself must construct/generate them (Semantic Processes). In the latter case, the set of possible models cannot be expressed as logical propositions of the symbols themselves. The logical propositions can evoke the semantic properties that we ground them in externally, which may make them feel like Semantic Processes to us. What I describe as a Semantic Process is what does that grounding, and which therefore can form those propositions. Like all the other definitions in this section, this Semantic Information definition formalizes a product of semantics, not the process.

what micro and macrostates to form and at what level. Different micro and macrostate choices can lead to different amounts of measured Information for the same message.

Theory 2: Semantic Information as Truthful Logical Assertions

The notion of Semantics being associated with truths has been a core assumption in many Semantic Information Theories (e.g., (Floridi 2019, Dretske 1981, Grice 1989)). One motivation is that Shannon Information does not refer to truths, and Semantics has been defined based on necessary and sufficient conditions in Logic (Tarski 1935) and Natural Language (Davidson 1967). A truth-based measure is in part justified by addressing the Bar-Hillel-Carnap Paradox (Bar-Hillel and Carnap 1953), where semantic information is maximal when one receives contradictory information. (Floridi 2005) resolves this counterintuitive consequence for measuring semantic information with respect to the probability of occurrence in a world by requiring that only true Semantic Information is informative.⁴⁷⁶ Alternatively, truth can be replaced by a receiver's belief: Something with 0% probability will not be believed or true. Measuring semantics based on truth conditions would also make some sense for guiding AI to tend to believe what is actually true.

Why Truth Conditioning does not cover Grounding: Constraining semantics to the truth of assertions is intuitive in many cases but does not address how symbols become understood as representing the world. This is fine as a Shannon-like model about facts of the world.⁴⁷⁷ However, the original driver of developing a Semantic Information Theory was that Shannon Information did not address meaning. A truth conditional theory does not in itself address meaning if it relies on someone, who already understands what symbols will signify, to first determine how those symbols will be associated with a probability distribution that reflects uncertainty about a world. In other words, grounding determines how each fact about a world is understood (i.e., what processes they activate for how to evaluate each message as a fact with respect to a Knowledge System). Nor is this resolved by building a Knowledge System using the same symbols. Such a theory provides a symbolic construct of what grounding might produce, but any such model is just as meaningless alone as any Shannon Information model. It can only evoke meaning in us who can ground the symbols. The distinction is consistent with the experience that the same message feels grounded whether I know it to be true or false. The richness of semantic experience for a message has no fundamental dependence on its objective or perceived veridicality. Again, there is utility in having an external measure for Semantic Information conditioned on the truth of assertions, but this is not the same as measuring the semantics from

⁴⁷⁶ Note that in Shannon Information, an impossible observation should also have the highest information. I could speculate that the reason this is only considered a problem for Semantic Information is due to the assumption that an impossible Shannon-level message cannot be observed. However, it seems the real difference is not Shannon vs. Semantic, but Objective Truth vs. Belief. Truth cannot be contradictory and thus no *true* contradictory message is possible. However, one can believe they receive contradictory assertions about some domain. Thus, a particular symbol sequence may not be possible at the Shannon or Semantic Information level microstate, but a transmitter could believe it is possible or even that this is what was transmitted (e.g., due to message corruption, measure/decoding error, etc.). A receiver would also be unable to receive a Semantic Information level *objective* fact about the world that is contradictory.

⁴⁷⁷ Again, Shannon Information can be measured for distributions over any symbol set (e.g., letter entropy, world entropy, fact entropy).

the perspective of the receiver itself. If semantics is understood with respect to perception (not conception), and thus subsymbolic, then this helps understand why a Truth-Conditional definition for Semantics does not work for Grounding as Semantics. Veridicality is only a property of Concepts. While Percepts just “are”, a truth property is directed at some symbol or set of symbols.⁴⁷⁸

Note that, while certain Semantic Information Theories may have a problem with the Bar-Hillel Carnap Paradox, grounding itself might not. A self-contradictory statement might be impossible to ground in a stable manner, which is consistent with how paradoxes are experienced. They oscillate as if they can never settle whenever a receiver tries to interpret them. One can create two incomplete and exclusive interpretations, but a unified interpretation is elusive. Only when interpreted with a different model, such as one *about* the soundness of the assertion(s), can stable grounding be achieved. That is, there are reference frames for which contradictions/paradoxes, redundancies, etc. are understood *as* contradictions/paradoxes, redundancies, etc.⁴⁷⁹ Paradoxes may be uninterpretable for the macrostate where the paradox occurs, meaning there is minimal semantic information at the level of this reference frame. However, at another level, the message can be interpreted as being a paradox or self-contradiction, and by being at this different macrostate, humans can resolve that it is a paradox, which is semantically informative.⁴⁸⁰ There is no paradox or contradiction at this level, since it is true that the statement is indeed paradoxical or contradictory.

While I avoid placing *objective* truth as a necessary criterion for Semantic Information, there are clearly activations that a receiver would treat as beliefs, sometimes with associated confidence in the truth of those beliefs. Learning that touching fire is painful conditions the receiver to act more effectively in the future. It is therefore important to consider how a fact can be associated with the receiver’s belief of its veridicality.⁴⁸¹

⁴⁷⁸ Note that with any learning models, falsehoods, mistakes, and contradictions ideally become increasingly unlikely as evidence accumulates. Beliefs should be capable of weighing evidence. Of course, this is not always the case, and may require many sophisticated ways to weigh evidence, but we do learn from the world because the world tends to bias messages towards reliable associations, allowing us to test assertions. In other words, we may not assume some objective truth about data with respect to some domain, but since the data generated comes from the world, it will continuously influence the model. Note also that, since Semantics is constrained by past effectiveness, significance could be measured separately with Effective Information. In this way, truth, when it correlates with effectiveness, is accounted for indirectly.

⁴⁷⁹ Another example is that the message “nsflnsx” has high Shannon surprisal given a model over letter sequences of English text, but a human’s Semantic Processes could also select a macrostate with one of the possible microstates being “nonsense”, which would have a different Shannon surprisal. There could also be some macrostate selected that includes what caused the transmitter to transmit the message.

⁴⁸⁰ *Research: An additional component of reasoning in a Conceptual Space should be knowing when one does not understand something or is missing certain information to reach a desired concept state or series of states. Determine how humans know they are likely missing information and therefore need to interact with the world to learn more (e.g., web search, asking someone, picking something up, looking around the corner).*

⁴⁸¹ *Research: How is unclear. One possibility could be that there is something like a veridicality dimension in the Conceptual Space for some activated paths. This way, a single symbol replacement (e.g., school → home) between two messages could have a different value along this veridicality dimension. The higher the value a message activated has along this dimension, the higher the receiver should condition future reasoning and decisions on the associated*

Theory 3: Semantic Information as Intended Meaning

(Weaver 1953) described the semantic problem as, “*How precisely do the transmitted symbols convey the desired meaning?*”. Since then, others have also considered how to measure Semantics with respect to intersubjectivity (see (Gärdenfors 2014) pg. 91 for references). Interpreting what a transmitter meant is not different in kind to gaining information about some world. On the receiver side, it can involve inferring based on what is more semantically probable. However, there can be a difference in the macrostates because determining what a transmitter said from a set of possibilities differs from determining how those interpretations change our understanding of the world (i.e., the microstate about the way that world really is). I therefore consider this separately. Accurate semantic decoding with respect to the meaning a transmitter desired is an important practical problem. Indeed, as (Gärdenfors 2014) points out, intersubjectivity is integral to Semantics and Cognitive Linguistics.

Why Intended Meaning does not cover grounding: It is straightforward to see that the same issues with the above theories occur when Semantic Information is applied to determine a transmitter’s intended meaning. The intended meaning of a message is simply a different domain from which to form possible micro and macrostates. Consider the following: The transmitter knows that, if a burglar is in the house while I am away to call me and say, “Rex is not a tyrannosaurus” as an alert signal. In this case, the use of the term is very informative and different than the “semantic” microstates for which it was applied earlier. But again, the macrostate this is applied to (the probability of the microstate that there is a burglar in the house) can be measured with Shannon-like Information. The actual Semantics is in forming micro and macrostates associated with the probability that a burglar is in the house, and not the micro and macrostate of what type of animal Rex is or the micro and macrostate associated with word sequence probabilities.

Again, it should be evident upon self-examination that the transmitter’s intention does not fully determine the quantity of Grounding evoked in the receiver. While not understanding what a transmitter means can reduce grounding, misinterpretations can be just as semantically rich as correct ones. More generally, the measure for grounding for the receiver’s model is not strictly relative to the transmitter’s model.⁴⁸² That is, if it is the meaning evoked by messages that we are after for a definition of Semantics, then there is some measure that does not need to reference some external agent’s intention or some objective truth.

To be sure, grounding in the mind of an individual is heavily influenced through interactions, especially with others as we seek common ground in communication. Grounding itself likely evolved, at least in part, for this purpose. **Grounding should make available, for example, a measure of similarity for symbols, and this could be used to measure similarity between a**

fact being true (e.g., It is New Year’s Day, and it would be odd for the transmitter to be at school right now). This veridicality dimension could be binary or include a continuous range for level of certainty in the fact being true. I am skeptical that brain’s emerge dimensions that so neatly fall into explicit qualities, but perhaps something approximate might be found.

⁴⁸² After all, when a person is isolated, even with full sensory deprivation, they continue to have thoughts with semantic properties.

transmitter and receiver's interpretation of a message. In other words, a fundamental semantic measure of what is evoked can enable use of a Shannon-like Semantic Information measure by comparing two models to find what (Gärdenfors 2014) refers to as 'Common Ground'. If *Person A* says 'dog' to *Person B*, then they understand each other to the extent their respective percepts overlap. There is some semantic distance between their grounding.⁴⁸³ Comprehension's role in finding common ground might be thought of as an attractor that helps minimize this relative distance between *person A* and *B*'s concepts by a series of transforms (iterated through discussion).^{484,485}

Measuring and optimizing for encoding and decoding messages to convey the transmitter's intended meaning will be important in AGI. However, it is not the only possible measure of Semantic Information. For example, it does not capture the creativity generated from the evolution of thoughts occurring across communication events. There may be cases where we would want Semantic Information to account for *new* meanings, which could occur from misinterpretation, or more commonly, where the value of a discussion is not just understanding one another but arriving at insights that neither agent had prior to the discussion.

Conclusions / Discussion

Is a thought about a unicorn as grounded as one of a horse? What about a factual perception of a real horse far away vs. imagining that horse in front of me when it is not actually there? All of these concepts seem semantic/grounded. To point to some objective factor seems arbitrary⁴⁸⁶; a unicorn can be associated with rich and meaningful visions. Each of the definitions above scopes Semantic Information to one specific reference frame (e.g., what a transmitter intended to transmit, what is the case in some world) but none seem to measure the actual role Semantics has in processing messages.⁴⁸⁷ There is therefore no need to claim one is the "true" measure of

⁴⁸³ *Research: However, it is difficult to determine how to measure it directly since the two persons' vector spaces differ. There are some Theory of Mind-processes (Goldman 2012) coupled with interactions between agents that aim to settle on some level of confidence that they are semantically in sync. This process should be investigated for AI interactions, both with humans and other AI agents. In general, Theory of Mind research should start with a foundation of Semantic Processes and Conceptual Spaces to implicitly ground beliefs about others' mental processes.*

⁴⁸⁴ *Research: This involves a lot sampling and inference because we cannot observe or directly interpret the internal models of other agents. This can lead to misinterpretations by the communicating agents, but humans have likely evolved/learning a variety of heuristics to improve the chances of reaching Common Ground. Identifying these could be valuable to implement in AI. Using AI to identify different barriers to reaching Common Ground may in turn be useful for improving social challenges in human societies. This mode of misinterpretation may also be interesting to consider with respect to cultural and knowledge differences. Greater differences in vector spaces and learned transforms make it harder to minimize semantic distances.*

⁴⁸⁵ *Research: Consider how this might be measured as a free energy minimization (Friston 2010) process between the two vector spaces, where minimum is always estimated, since we cannot directly observe the other agent's "vector space".*

⁴⁸⁶ A guess on my part is that one of the motivations for measuring Semantic Information against some truth in the world is that it seemingly bounds the set of possibilities to measure against. Otherwise, how could one explicitly account for the number of possible messages and possible models? This issue is avoided by non-counterfactual measures, such as Semantic Processes.

⁴⁸⁷ I have not addressed every existing theory on Semantic Information. For example, a more recent approach (Kolchinsky and Wolpert 2018) associates Semantic Information with the impact of interventions with respect to

Semantic Information, since there is a more fundamental process to determine the micro and macrostates that set the reference frame.⁴⁸⁸ While these Semantic Information measures demonstrate different magnitudes than Shannon Information and from each other, given the same message, this can be explained by a change in grounding associated with the formation of different micro and macrostates. In this sense, the process that forms the micro and macrostates (i.e., constructs the symbol indices) is equivalent to interpreting messages. In contrast, the Semantic Information theories above all *assume* the interpretation. There is no accounting for how micro and macrostates are selected, which I claim is the core function of Semantics. Without this grounding, the semantic qualities they propose to measure rely on what is evoked in us as external observers. To my knowledge, no AI can be built that can make effective use of the types of semantic information these theories offer to understand messages. They would require either grounding applied through another formalization or human modelers to predefine all the mappings for terms, therefore bypassing any of the actual need for the modeler to interpret them. The latter is brittle and not scalable.

Since there are no fundamental Semantic Information Theories, it might make sense to consider having two broad categories of semantic information measures: objective and subjective, where the former is measured with respect to some external objective truth and the latter against the receiver's beliefs. Indeed, for some AI models, one may want to measure both to be able to optimize them to converge.

While all these definitions can coexist for measuring different types of Semantic Information, I tentatively claim that Grounding should have a single definition. Even if externally measured, it describes an actual process rather than a counterfactual comparison against externally defined possibilities. Neither objective facts nor transmitter intention have a direct role in measuring Grounding. Even social consensus on specificity is not an objective external measure for concepts. If a message evokes something more complex in their Knowledge System upon receiving the message “mammal” than they do for “dog”, then why would this not be more semantic? The measure for grounding must rely solely on the properties evoked in a receiver's Knowledge System given a message, and this can differ depending on the receiver.

Finally, if a Semantic Information theory is nonzero but the system measured is incapable of grounding, then it is absurd to claim the model itself has semantic understanding in any way (in contrast to some external semantic observer). This is why Grounding is a key measure that enables measures for Semantic Information for systems themselves. Thus, Grounding should not

survival of the receiver. However, this seems more like what Weaver would have meant by Significant or Effective Information. (Weinberger 2002) also models meaning as “pragmatic information”, using the assumption that meaning is derived from utility of information. Effectiveness and Semantics shape each other, but I believe it is more appropriate to have a separate measure for effectiveness, which could be used to optimize the architecture and updating method of the receiver's knowledge system and how semantics encodes and decodes messages, therefore shaping semantics, but as its own thing. In this way, Meaning is itself a heuristic adapted for its utility, and its properties are shaped by their utility, but the properties associated with Meaning would not be constituted by utility.

⁴⁸⁸ *Research: Consider if meaning tends to be directed towards a reference frame where grounding is stable via some sort of energy minimization. If so, then the relationship between the initial but unstable state and the resulting state as a lower energy state, along with underlying force dynamics, is worth further study.*

be seen as contradicting these measures and can, in fact, enable them to be measured effectively for AI.

Semantic Compression

Introduction

Since symbols are derived from some subsymbolic base, basing Semantics on just symbols would be to focus on a product of Semantics rather than Semantics itself. Moreover, this product is often, in a sense, extremely lossy. Cognitive subsymbolic processes may be difficult to accurately or reliably evoke in ourselves or others through the medium of symbols. However, Semantic Processes cannot be communicated without the use of symbol sequences, like the ones you are currently reading.⁴⁸⁹ The evolutionary advantage of symbols is that they make communication possible.⁴⁹⁰ Symbols are how we evoke a common ground, through mediums that separate mental models, for the purpose of coordination. For some reason, they also seem to be used for internal communication (i.e., explicit thinking). **A key evolutionary advantage of Semantics relates to the ability to compress messages beyond a single symbolic level, unlike “Shannon-based” compression.**⁴⁹¹ At the grounding level, knowledge provides context used in decoding⁴⁹² a message into an “appropriate” macrostate that is updated based on that decoding. Thus, Semantic Processes play a necessary role in enabling Semantic Compression by forming macrostates and microstates that act to decode messages using existing knowledge. In contrast, Shannon-like Compression involves transforming a sequence of elements into typically smaller sequences but from within the same symbol set. The macrostates and microstates are already determined.

⁴⁸⁹ *Research:* It is worth mentioning that model forming and model updating might occur simultaneously. In other words, even if there is Formal Cause from the Semantic Processes to the Shannon Information updates, there may or may not be an Efficient Cause. If there is only a Formal Causal relationship, then Semantic vs. Shannon-like descriptions could be complementary perspectives of the process. *A useful study could be to determine under what conditions the Formal Cause occurs vs. when it is absent (i.e., no symbols are formed).*

⁴⁹⁰ I suppose we could in theory connect our brains directly. Direct neural communication could one day enable subsymbolic communication. However, it is unclear that this would omit the need for symbols, since we clearly use them for internal communication (e.g., thinking).

⁴⁹¹ *Research:* Semantic Compression is not the only benefit of generating symbols. Symbols enable something like “mental teleportation”. For example, a grounded perception of a lion can be evoked by sending the word ‘lion’ as a symbol. Being about a lion can evoke it without having to actually be in the presence of an actual lion. Moreover, the compositionality of symbols enables considering objects and/or events of interest that may be almost impossible to actually experience, such as visiting Saturn. Other advantages include the ability to plan without needing to have the actual experiences for which plans are generated and communication of objects/events that others would not likely be able to experience. I am not arguing that subsymbolic processes necessarily require symbols to evoke perceptions of phenomena outside of those directly sensed. However, we frequently use the power of symbols through communication to evoke such phenomena. ***We should determine whether they hold some advantage in doing so vs. purely subsymbolic processes. More generally, determine if/how internal explicit reasoning provides unique functionality over simply performing all mental operations over the subsymbolic knowledge space.***

⁴⁹² Whether we call this ‘decoding’ or ‘encoding’ depends on the perspective. For most of this chapter, which focuses on the receiver’s Semantic Processes, the message is what a transmitter encoded into symbols as evoked by their grounding, which would then be decoded by the Receiver’s Semantic Processes into their Knowledge System.

However, there is not much new to say about this role that has not already been said above, beyond its key role in both increasing knowledge and encoding and decoding of messages using that knowledge. Instead, I will mostly consider Semantic Compression at the symbolic level, which can use existing Semantic Information Theories. These can be contrasted with standard Shannon Information-based Compression methods, where symbols can be differentiated based on “basic” symbols (e.g., letters in the alphabet). Grounding is only discussed where there is something to say regarding Semantic Compression that has not already been discussed.

The basic idea is that relevant knowledge can be used to capture information from messages that is not directly encoded in the messages themselves (or more accurately, knowledge that goes beyond the standard mapping available to alphabetic sets, such as number systems and ASCII). Less knowledge to interpret messages generally means larger messages must be sent. Using dice rolls as an example, if someone does not know how to count, then that must be explained in messages along with the dice role.⁴⁹³ Without also transmitting information on how to count, the intended semantics are missed, not due to noise or channel capacity, but because the transmitter did not accurately account for the recipient’s limited knowledge to properly ground and evoke the desired symbols.⁴⁹⁴ Thus, **knowledge can reduce the size of the message required to extract the same amount of Shannon Information**. By combining what a receiver and transmitter both already know with the message, additional redundancies in the message can be removed prior to encoding them to transmit. What the message is meant to convey is decoded by the receiver “completing”⁴⁹⁵ it using existing knowledge. For example, knowing what ‘the refrigerator is broken’ “means” removes the need to formally communicate features and functions of objects in messages that the transmitter may desire to evoke in the receiver (e.g., refrigerators keep food cool, it is safer to store food at cold temperatures, one can pay for services to fix it). For Semantic Compression, **common knowledge is like a decoding key, where a microstate can be the transmitter’s model, for which the message is an approximate encoding. For the transmitter, the encoding applied is driven to evoke a certain Grounding in the receiver. Thus, for the transmitter, the micro and macrostate are associated with the receiver’s model.**⁴⁹⁶

Before moving on to details about Semantic Compression, I come back to Grounding temporally to make a claim relating Compression and Emergence: **Emergence itself may be functionally attributed to the semantic compression it provides. There is loss in this process, associated with irreversibility. This is why emergent level properties cannot be precisely**

⁴⁹³ Instructions for counting instructions will need to be sent with every roll if the receiver has no memory faculties.

⁴⁹⁴ *Research: It would be interesting to study how much this explains why different opinions are difficult to resolve. If two people have different knowledge, but it is difficult to detect the differences, then this could affect how they ground new information.*

⁴⁹⁵ “Completing” is put in scare quotes because it is not necessarily true, or perhaps even common, that existing knowledge literally fills in the blanks with missing symbols. More often there is simply some implicit semantic understanding evoked. Thus, Semantic Processes (rather than Semantic Information) may often be sufficient to characterize the processes. However, there is nothing obvious to add here that has not already been said in earlier sections. There is plenty to say about the role for Semantic Information.

⁴⁹⁶ Of course, without Semantic Processes, there is no process to activate one of many possible models.

mapped back to the lower levels underlying their encoding and decoding. However, biological intelligence emerges grounded concepts because they tend to help cope with the world better than expending the resources required to model the world at lower levels.⁴⁹⁷ Semantic Compression, at its core implicit level, provides a way to emerge higher level concepts from lower-level ones. Since this model-to-message encoding is irreversible, there is no explicit map to decode back to the transmitter model's exact properties encoded.^{498,499,500} As heuristics, the ultimate value of such emergent compressed concepts would be assessed via some Effectiveness measure (e.g., effectiveness in reaching implicit or explicit goals).^{501,502,503}

Message Completion and Abstraction

Semantic Compression can be applied in at least two related ways. First, Semantic Information leverages existing knowledge to **complete messages**.⁵⁰⁴ That is, semantics not only activates models given messages but can use those models generatively to complete messages. Messages often consist of multiple symbols, the grounding of which may be influenced by the grounding of previous messages (and their associated updates to the Knowledge System). As the symbols are received from a message, one interpretation may become increasingly probable.⁵⁰⁵ At some point during the receipt of a message, the model may become so probable that the full message is not needed. When the message completion is performed using symbol sequence probabilities, standard Shannon Compression can be applied. Semantic Compression makes use of other domain knowledge (i.e., knowledge about how symbols relate to things in the world), as discussed in the Semantic Information Theories section above. For example, one can use existing knowledge to know that a text about someone who just had a baby that says, "g or b"? is a

⁴⁹⁷ We would require more capacity than is available in the Universe for many lower-level environments, at least to build full explicit models of low-level variables and their possible interactions.

⁴⁹⁸ *Research: Semantic loss is not necessarily the same as Shannon-based lossy compression, although potential relations are worth further consideration.*

⁴⁹⁹ Note: Just as for Shannon Information, amount of Semantic Information would not be changed by lossless message compression. Thus, amount of Semantic Information should not be defined as proportional to amount of message compression. Semantic Information, aggregated over messages, will increase the amount of compression possible, but this is more of a measure of knowledge.

⁵⁰⁰ *Research: Does explicit thinking (i.e., intra-agent communication) involve a full encoding and decoding processes? Consider how the information loss involved could be measured.*

⁵⁰¹ *Research: Consider if/how semantic information could be related to the joint compression the message enables across a data network (i.e., how much bandwidth is reduced across the network by adding and enabling use of shared knowledge across nodes). As a simple example, if a symbol from a message relates two separate symbols in the network, there should be some way to quantify the semantic information provided by the change in the network, proportional to the extent that future messages in the network no longer need to include both symbols. Such a measure could be useful to optimize such networks and potential data sources.*

⁵⁰² *Research: Consider if there is a way to approximately measure loss through the process of forming explicit symbols from semantic processes (e.g., measuring differences in messages or neural states given extensively repeated generate/receive events).*

⁵⁰³ This all relates to an ontological claim: As far as I can tell, all explicit concepts are lossy compressions of a subsymbolic reality.

⁵⁰⁴ This is similar to a Hopfield Network (Hopfield 1982).

⁵⁰⁵ Or one interpretation becomes the dominant attractor to complete the message.

compression of “girl or boy?”. The conditional probability based on the letter or word sequences alone would be unlikely to be able to decode the message as “girl or boy”. This can reduce message size since shared knowledge can be used to complete it by evoking the same model formation that the full message would have evoked.

A second form of Semantic Compression is **Message Abstraction**. Like Message Completion, Message Abstraction also compresses message sizes by forming models that ground the message symbols or symbol sequences, but to emerge abstracted concepts. Abstractions can make assertions across sets of individual lower-level semantic phenomena, which is useful when the impacts apply to each. At the subsymbolic level, abstraction enables encoding and decoding within a hypervolume of conceptual space that presumably covers the subsets relevant to information communicated. Existing knowledge allows the formation of macrostates at the appropriate level without needing to share all the information about the individual members and their individual properties. For example, most English-speaking receivers who receive ‘mammal’ in a message can use their existing knowledge to decode all the associated properties of mammals and apply the information evoked from the message to all mammalian species.⁵⁰⁶ There is no need to provide separate messages for dogs, whales, kangaroos, etc.). Note that these decodings are often implicit. Clearly, no human’s mind actively considers each mammal type and updates its properties given a message. Instead, it can be thought of as information potentialized by modifications made to the underlying Conceptual Space. What future messages will evoke has been modified. Similarly, the notion of “completing” or “abstracting” messages is more accurately described as an implicit change to the receiver’s Conceptual Space that constitutes the experience of comprehension.

Of course, the grounding evoked for ‘mammal’ also differs from a list or graph of its members and properties. The semantic compression changes grounding properties.^{507,508} The role of Semantics in abstraction with novel properties relates it to Mental Chunking as Semantic Compression (Nassar, Helmers and Frank 2018).⁵⁰⁹ This compressed form enables more effective use, including for memory, communication, and inference, of mammal-related assertions. I further suggest that such compressions do more than simply bundle members into groups to efficiently share information. **They also can be used to efficiently avoid the notorious Frame Problem (Shanahan 2016) in AI through their grounding being shaped by Effective Processes.**

⁵⁰⁶ Again, this is a semantic process since ‘mammal’, just like any other symbol at any abstraction level, must be grounded (i.e., interpreted) to determine the model for which the information in the message will be actualized to update as knowledge.

⁵⁰⁷ That is, abstractions are not simply the aggregation or averaging of what their components share, as the abstract concepts may have novel properties.

⁵⁰⁸ *Research: Determine if there are systematic properties for how abstractions achieve novel grounding in Conceptual Spaces, as well as what contingencies in evolution, if any, led to any innate novel grounding properties for specific abstractions (that is, some abstract concepts may have been important enough to be distinctly shaped within our Conceptual Spaces in ways that are independent of their individual members).*

⁵⁰⁹ *Research: Formalize these Message Completion and Message Abstraction processes in Conceptual Spaces and then consider how Semantic Compression and Novelty can be measured. Consider how mental chunking of words in a sequence could be measured in a Semantic Vector Space (e.g., something like the difference in the sum grounding of the individual symbols vs. the grounding of the chunked subspace (or resultant vector)).*

Our groundings for classes of entities, events, etc. are incomplete, but in a different way than an ERM would be incomplete. Instead of attempting to explicitly structure the literally boundless types of properties that could make up such classes, our compressed concepts have implicit practical relationships to explicit properties. For example, danger, size, and speed may be most important for mammals in certain contexts. How mammals are grounded in other contexts will differ. A model that attempts to explicitly specify all of these relationships would be intractable. Effectiveness influences the grounding of abstractions to evoke properties that are most relevant to (i.e., effective in) different contexts.

Statistics itself is Compression. For example, linear regression compresses data down to a y-intercept and slope equation. When performed at the level of features processed, statistics is a form of Shannon, Algorithmic or other Information-based Compression. Scientific theories/laws might also be considered compressions of the phenomena they describe. They often provide Semantic Compression, including use of abstraction for emergent concepts. As always, it is the grounding that these compressions as concepts evoke that provide their meaning. There are potential implications for understanding these processes as compression. **AGI will need cognitive tools to cope with the world, including those we have adapted/developed ourselves. Conceptualization have been vital to our intellectual and technological advances.** For example, the questions we ask ourselves influence what can be statistically compressed. Consider that the Earth's orbit can be described using only the center of mass and distance to other masses. This is due to the emergent heuristic nature of not only the earth but the question we are asking. **Currently, the approach to AI is not so much to imbue AI with these compression capabilities, but to model the AI itself with them. Again, we see the distinction between the models we build vs. the modelers that AI must become.**⁵¹⁰

Many ANN architectures enable both message completion (Hopfield 1982) and feature abstraction. These include Autoencoders (Autoencoder n.d.) and Convolutional Neural Networks (Krizhevsky, Sutskever and Hinton 2017), where the number of parameters is reduced across layers to learn the patterns that can then be generalized at a higher abstraction level. Both message completion and abstraction, as Semantic Compression, also relate to the concepts of Information Inflation and Deflation (Haken and Portugali 2014).⁵¹¹ A receiver with a rich enough

⁵¹⁰ The relationship to Bias vs. Precision, Context Adaptation, and Causal Modeling will be discussed in a future manuscript.

⁵¹¹ There are purely symbolic ways to complete messages. Shannon (Shannon 1948) himself demonstrated how increasing the order of a Markov Chain can increasingly produce more sensible natural language text. Here, the variables are letters with transitions conditioned on what previous letters were generated. The 'order' is the number of previous letters used to condition the probability of the next letter in the sequence. Related studies have been done for human performance in completing text (Shannon 1951, Miller and Friedman 1957). This form of inference can be generalized to Conditional Shannon Entropy $H(X|Y)$, where Y is the context, and this context can be potentially interpreted as the grounding for X (the predicted symbol). Context Y can be as complex as necessary, such as capturing relationships among messages. Of course, I claim there are considerable differences in what Markov Chains vs. what humans are doing. Humans, who have trouble multiplying a couple 2-digit numbers, do not seem to be explicitly computing complex statistics. And yet Shannon found that human performance is nearly optimal. Conceptually speaking, humans must determine symbol indices using subsymbolic processes. As an implicit process, this avoids the need for computationally expensive statistical modeling at the level of the symbols

semantic vector space could evoke additional dimensions not “explicitly referenced”⁵¹² in the message itself, which is a form of association capabilities. For example, I could picture a hot coffee when receiving a “coffee” symbol from a message that is not accompanied by a “temperature” message element. This uses Information Inflation, which enables a transmitter to send a smaller message. Information Inflation might also be used when a higher level macrostate is ambiguous. Abstraction, on the other hand, deflates Information given the message, and thus decreases what the message theoretically could provide to the receiver.

Semantic Redundancy

When a message can be accurately⁵¹³ completed or an abstraction decoded by a receiver using existing knowledge, sending the complete or more granular message can be described as **semantically redundant**. Semantic Redundancy depends on the receiver’s existing knowledge and has the same costs and benefits as Shannon Redundancy in trading off message efficiency with the probability and impact of the receiver decoding something different than the transmitter intended.

Since transmitters do not have direct access to a receiver’s knowledge, they must predict what information is semantically redundant.⁵¹⁴ The receiver’s knowledge therefore determines the ideal minimal message size required to decode a given desired interpretation. In some cases, it may make sense to think of messages over a lifetime, rather than single episodes. Education can in part be thought of as the application of related messages with the aim of increasing semantic compressibility of future messages to allow agents to cope with the world in ways that produce social value. Expertise involves not needing to receive full explanations already learned.^{515,516} Since the transmitter must infer what receivers know, Semantic Redundancy is a tradeoff with Semantic Compression to determine the potential for semantic error (i.e., misinterpretation). Consider the text message: “Edgar is coming over and wants to see you.” If I infer that the receiver of this message will know that I am referring to our mutual friend Edgar Smith, then I expect that adding “Smith” to the message would be semantically redundant.

themselves. With Shannon’s Markov Chain, the symbol indices are always predetermined and the same would need to occur for a Markov Chain intended to capture the full context required to perform at human level. Explicitly constructing and calculating over necessary “context” is not only an issue of combinatorial explosion. It would need to capture a variety of implicit variables as context.

⁵¹² “explicitly referenced” is an inaccurate portrayal here since messages themselves do not intrinsically reference anything. It is more accurate to say that, counterfactually speaking, certain grounding, possibly including specific qualities, is evoked in the receiver that, without relevant knowledge, would require additional message elements to be transmitted.

⁵¹³ “accurately” here meaning as intended by the transmitter.

⁵¹⁴ They must also infer whether the receiver did decode the message as intended, which might be assumed given experience and/or assessed based on subsequent communication.

⁵¹⁵ This is an oversimplification since there are physical and information capacity properties that likely constrain how messages must be split up over time. For example, humans cannot simply be provided one long message explaining everything they need to know to architect a building immediately before being presented the task.

⁵¹⁶ *Research: What about for intra-agent communication? Can our minds always select maximally compressed messages for ourselves? Consider how one could measure the difference in semantic compression between inter-agent and inter-agent communication (e.g., via self-reporting).*

Replacing “Edgar” with “He” is even more compressed but is unlikely to be correctly interpreted by the receiver without additional context. Thus, “He” would lead to semantic error, resulting in the need for additional messages to clarify. “Edgar” is therefore a tradeoff between Semantic Compression with potential Semantic Redundancy.⁵¹⁷

The above example describes the inference process of what a receiver knows in an explicit form. In reality, it would be intractable for agents to somehow compute probabilities over all explicit concepts that a given receiver may know, including how those concepts relate to more or less abstract concepts that could be encoded. **The ability to predict/infer what will be evoked in receiver’s given messages, is largely an implicit process that takes advantage of shared experiences and the fact that agent models are likely similar given our ancestry and similar environments. If an agent wants to evoke something semantically in another agent, they can infer what that agent likely knows based on similarities and differences in experiences, and they can infer what message symbols tend to evoke the model property changes in themselves that they want the message to evoke in the receiver agent and transmit those, with some expected loss tradeoff. This may not always be necessary but is a more plausible explanation for how Theory of Mind⁵¹⁸ works than, say, as a probabilistic ERM representing explicit beliefs. If I am correct about the criteria for Semantic Processes (i.e., Grounding), then they are required if we want agents to have Theory of Mind, or more generally to evolve/learn any grounded prediction/inference processes.**⁵¹⁹ For the receiver, the microstate can be the interpretation the transmitter is trying to convey and the macrostate is the set of possibilities for this interpretation given the message received. For the transmitter, there is a macrostate that includes the possible grounding a candidate message would evoke in the receiver. In a sense, this is a supra-macrostate consisting of possible interpretations of the receiver’s possible interpretations, which represents at least one possible form for Theory of Mind.^{520,521}

Semantic Encoding

How can optimal Semantic Compression be determined? At the explicit level, if one had complete access to what a receiver already knows and how they would decode any message

⁵¹⁷ Winograd Schemas (Ackerman 2014) are great examples of Semantic Compression, where the antecedents for pronouns are ambiguous without applicable knowledge.

⁵¹⁸ Theory of Mind refers to the ability for agents to attribute certain mental states to others and to oneself (Goldman 2012). Ability to infer what others might think or feel given certain events or actions is important for social functioning. Theory of Mind is considered a higher cognitive capacity and an open research area in AI.

⁵¹⁹ Note that this creates another challenge for AGI; without sharing the evolutionary history or model structure, how can we infer what they know vs. what we know? Even in the case where communicating agents are evolving together, we as humans will be challenged to understand their semantic processes. Semantic Specificity, as a magnitude, is insufficient to know if the processes evoked are semantically similar without direct access across agent Conceptual Spaces. Potential paths forward, including ways to approximate the qualities of grounding, are provided later.

⁵²⁰ The receiver could also attempt to model probabilities over different possible assumptions the transmitter has about the receiver as a macrostate. Thus, Theory of Mind can in principle become as arbitrarily meta as can be modeled. In practice, this does not seem to go beyond a couple orders for humans before comprehension reaches its limit.

⁵²¹ *Research: Consider if Conceptions, as Perceptions of Perceptions, are also Compressions of Compressions.*

given that knowledge, then one could in theory⁵²² figure out the minimal message that could be sent to evoke the decoding desired. This would be the optimal **Semantic Coding**. A transmitter may expend mental resources to search for some level of clarity and conciseness.⁵²³

However, humans do not necessarily attempt to reach optimal Semantic Coding for a single message. It is often more efficient in the long run to interact with receivers to reassess compression needs. Doing so often improves the Semantic Coding results by learning more about what the receiver understands to inform how a message will be encoded to transmit to them. This is often achieved through iteratively finding common ground for smaller messages. Breaking up messages reduces the complexity and uncertainty that a transmitter must account for to achieve accurate inferences for the set of possible receiver interpretations for candidate messages to transmit.⁵²⁴ Thus, interactions, and possibly even expectations for how those interactions may go, can affect what semantic coding is optimal.

What this all suggests is that Semantic Coding is something that can be formalized, but there are practical challenges to measuring it that must be overcome. My hope is that this is seen as an opportunity to develop approximation methods as well as theory.⁵²⁵ The benefit of any such methods is considerable for achieving much greater message compression than current Shannon-based methods and without needing transmitters to first explicitly determine what receivers do or do not already know.⁵²⁶

Agent Coordination

Communication is, crudely speaking, a path dependent function of concept transformations. That is, messages within a dialog typically depend on previous messages received and transmitted. Semantics are often involved in a multi-step process whereby initial models provoke subsequent ones based on grounding given new information through message(s).⁵²⁷ (Gärdenfors

⁵²² This could require unlimited sampling with the model.

⁵²³ *Research: Since we do not have access to the full Knowledge System of potential receivers, a semantic coding measure must somehow consider potentially relevant past information and additional context to predict probability of possible knowledge systems and how they would be used by a given receiver. As far as I know, no such general semantic coding measure has been formalized for the common case of incomplete information. Also determine how Semantic Coding can take the subsymbolic processes into account.*

⁵²⁴ This is because, through interactions, particularly those aimed at the receiver reaching some ultimate correct decoding, the transmitter learns more about what the receiver knows.

⁵²⁵ *Research: For example, what is the maximum amount of semantic compression possible? This might be something like a unique symbol for each possible distinct interpretation that could ever be shared, and the unique number would be the same across all messages that have a common interpretation. However, this does not address the fundamental open-ended nature of knowledge. It seems that a theory must address semantic compression as a generative construct. Other Questions include: What must be the same across Knowledge Systems to achieve intended decoding? What is the optimal amount of semantic compression given energy costs and probability of correct receiver interpretations (assuming effectiveness could be controlled for)?*

⁵²⁶ *Research: One option to consider might be an algorithm trained to predict the smallest message that evokes the same grounding in a receiver given the receiver's inferred model parameters and current state as context (e.g., activated vectors given message effects in receivers' vector spaces (i.e., knowledge) to use for distance as prediction error). This requires access to a variety of different receivers as training data.*

⁵²⁷ *Research: In this way, perhaps Semantics can potentially discover patterns in messages across macrostates, although future studies are required to determine this. Consider what this entails using a Conceptual Space.*

2014) has shown how this can work by converging to a fixed point through back-and-forth communication.⁵²⁸ One can in theory, therefore, identify a state or series of geometric shapes and transformations in Conceptual Space that a transmitter wants to evoke in receivers to move to some desired goal. Goals can include evoking the same state encoded by the transmitter or a different state that the transmitter expects a message to evoke in the receiver(s). Similarly, when thinking about a problem, it seems the mind has some notion of the conceptual meaning it wants to grasp, possibly as a series of implicit criteria to satisfy. The thinking process might then be described as seeking to evoke a form and location (or series of forms and locations) in the Conceptual Space that satisfies those criteria. Conceptual Spaces enable open-ended analogy making, which can be used to help each other learn what the other(s) mean by concepts communicated. Any model can have words mapped to models, including ERM as Knowledge Systems. However, the discrete nature of these requires manually building in each word or word combination. AGI-level Analogy Making is infeasible without the underlying metric space.⁵²⁹ This process should often result in new emergent concepts. In a way, then, this process for seeking certain grounding states involves sampling within one's own Conceptual Space to discover new (emergent) forms that satisfy certain implicit or explicit criteria.⁵³⁰ This is the creative process, particularly when constrained by Effective Processes to produce useful results.⁵³¹ **Understanding how Semantic Compression is optimized given common knowledge seems critical to advancing AGI not only to communicate at lower bandwidths / energy costs, but in gaining the ability to be creative.**

⁵²⁸ *Research: (Gärdenfors 2014) claims that Brouwer's Fixed Point Theorem (Brouwer 1911) and inter-agent interactions can demonstrate use of a common ground as a mapping between two Conceptual Spaces. Determine how and when symbols and symbol sequences would be expected to map to concepts in a similar position for both agents' metric spaces.*

⁵²⁹ *Research: Consider whether a benefit for generating symbols is using them as indices to retrieve certain semantic concepts. This may be an efficient means for memory retrieval. Given that inter-communication necessarily involves symbolic messages, it may be that intra-communication reuses this process. Perhaps this is an efficient way to generate externally what we were previously thinking internally; in many cases, it is useful to rehearse and hear back what we plan to say so we can hear it how someone else would (i.e., as if someone was saying it to us). This could produce higher semantic quality of transmission by understanding how we interpret the symbols we generate (thus acting as part of a planning process). Also consider the association of this possible function with Reasoning as a means of Justification (Mercier and Sperber 2017). Also consider if this claim relates to how (explicit) thinking is often in 2nd person form (as if making arguments to someone else). I do not at this stage feel strongly about these claims, but it is worth further study, including integrating understanding beyond natural language to account for the role of vision, etc. in cognitive processes.*

⁵³⁰ Criteria do not always need to be explicitly defined. There may be cases where they are not even implicitly predetermined by the agents. One might be satisfied with where thoughts lead without first setting satisfaction criteria. This is a challenge for AI with predetermined objectives that are too narrow to achieve general intelligence behaviors.

⁵³¹ *Research: Consider if maximizing Semantic Specificity or Semantic Information can sometimes be analogous to greedy exploitation with some stochasticity to enable exploration. This stochasticity could be in the form of noise generation, perhaps triggered by semantic lossiness of inter- and intra-communication.*

Some Alternatives for Characterizing Grounding or Semantic Information

In this section, I describe several alternatives to the way I have so far characterized Semantics. Not all of these are exclusive to claims above. Some of these are more thought out than others. I leave it to future research to integrate them or identify where the mistakes lie.⁵³²

Semantics as Actualization of Information: Another measure of possible interest is to develop a proxy measure based on its duality with Shannon Information. The rationale is that Semantics, by forming a model, actualizes Shannon Information given a message. Specifically, we can measure a proxy for Grounding based on what the model formed actualizes in terms of Shannon Information. While a potentially useful measure in its own right, I do not see this as a valid alternative form measuring grounding magnitude. It would make Grounding and Shannon Information too correlated, given convincing examples by (Bao, et al. 2011) that they are not always equal for the same message (i.e., one can be larger than the other depending on the message). Grounding as I have described it earlier does relate to Shannon Information in that forming the micro and macrostates is required to measure a Shannon update to the macrostate. However, this dependence does not entail any strict correlation in their magnitudes.

Semantic Information = Lower-Level Shannon Information: One possible way to measure Semantic Information is as a conceptualized form of model formation. That is, if the model formation process is predefined as symbols (i.e., modeling different possible indices to select for a higher level macrostate), then one can symbolically capture dependencies between levels. This creates a nesting of lower levels measured as Semantic Information that each create a Shannon-like model at the next level. Table 1 provides an example of this nesting. Semantics determines that 'a' is a letter, and thus should be applied to the Shannon Model that is measuring probability of subsequent letters. At the next level, where the Shannon Model is measuring the probability of words, Semantics must be selecting the set of words for the indices, and so must recognize that a symbol received is a word. This does not capture the essence of Semantic Processes / Grounding because something must determine the symbols at the "Semantic" levels (i.e., the index of indices). Thus, while loosely related to Grounding, the measure is no more foundational than the other Semantic Information Theories described above.

Table 1. Examples demonstrating nested Semantic and Shannon Information. $P()$ = Probability().

Symbol Set	Semantic Information (I_{Sem})	Shannon Information (I_{Sha})
{a, b, ..., z}	I_{Sem} associated with $P(\text{Symbols are letters})$	I_{Sha} associated with $P(\text{letter})$
{study, activities, ..., goals}	I_{Sem} associated with $P(\text{Symbols are English words})$	I_{Sha} associated with $P(\text{word})$
{GPA = High, Experience=Good, ..., Criminal_Record=None}	I_{Sem} associated with $P(\text{Symbols are features associated with hiring quality})$	I_{Sha} associated with $P(\text{good hire})$

Semantic Information as impact on a receiver's beliefs: This measure is similar to Shannon Information, whereby a message is maximally informative given its effect on knowledge. This can

⁵³² Research: Determine how each of these hypotheses relates to definitions for inference methods and possible measures such as Conditional Semantic Information, etc. How does the process of reaching Common Ground relate to Mutual Semantic Information?

be quantified for a message, such as by the sum change in the receiver's Knowledge System and is already measured for many types of AI methods, including ERMs (e.g., Kullback-Leibler Divergence (Kullback and Leibler 1951) between a prior and posterior distribution given the message across a Belief Network). If performed over the explicit symbols representing semantic concepts, then this is not an intrinsic measure, but of course, neither are the other Semantic Information Theories described above.^{533,534} However, it can be measured on subsymbolic processes, such as during ANN training, where the relationship to Semantic Specificity is of further interest.

Semantics is fully Qualitative and therefore not Quantifiable: Semantics is intuitively associated with the intrinsic phenomenality of experience, which we are limited to measuring from a third-person view. The third-person perspective is subject to eliminative materialism. I associated intrinsic emergence (and percepts) with qualia, where only Concept-based properties are subject to eliminative materialism. If Semantics is not only associated with intrinsic experience but equivalent to it, then there is nothing we can directly measure. Semantics would simply be what is left unexplained when everything has been quantified. The only definition for Semantics would be qualitative changes in qualia evoked by messages relative to the current state. As the qualia cannot be directly measured, any indirect measures are Shannon-like approximations. This may be inevitable as we cannot communicate in anything but symbols, including the models and measures we design. The challenge thus comes down to the issue that a chain of symbols mapping to other symbols (e.g., words to letters to bits) always leave out the meaning that the symbols are meant to represent, except in what they evoke in us extrinsically. As discussed earlier in this chapter, I believe a way forward is to define 'subsymbolic' as anything outside a given model's symbol set from a model-centric reference point. The criteria put forward involves measuring semantics using geometry within metric spaces. While still indirect, the idea is to capture what one would expect to observe for a model if it did involve Semantic Processes.⁵³⁵ Ultimately, if Semantics is distinct from Shannon Information and not epiphenomenal, then we should be able to test its effect compared to purely symbolic models, although observing effects is not the same as being able to directly measure or model it.⁵³⁶

Effective Information

This section is shorter than the Semantic Information discussions, not because Effective Information is less important, but because I have not given it the same attention.⁵³⁷ As a

⁵³³ Such a measure seems to handle contradictions as well as a truth-based definition, since two exclusive equally weighted assertions cancel out their impact to the Knowledge System, creating no impact overall to the belief. However, as discussed earlier, there are other semantics evoked, such as understanding the assertions as contradictory.

⁵³⁴ Again, I do not claim this should replace any other measure for Semantic Information. What should be used depends on one's objective. For example, (Bao, et al. 2011) seems more useful for measuring Semantic Communication.

⁵³⁵ A loose metaphor would be like covering something invisible with a blanket.

⁵³⁶ *Research: Consider if/what causal modeling approaches might be used to isolate the effects.*

⁵³⁷ I have not given it the same attention because it seems relatively more straightforward. However, there is plenty more to do in order to develop and implement it in measures.

reminder, (Weaver 1953) identified a third area for information not covered by Shannon's formulation for Information Theory, which relates to the significance of the information on the receiver for messages. I will refer to this as 'Effective Information' or 'Effective Processes', depending on the nature of the measure, or 'Effectiveness' when referring to either one. A description of Effectiveness should capture the claims in earlier chapters, particularly in the Performative Systems chapter.

My core claim is that **the Effectiveness of a message for a system/agent is how much the message contributes to achieving some future state(s) that maps to some objective(s)**. Before explaining this approach to Effectiveness, it is useful to distinguish different measures that one may have in mind for Effective Information or Effective Processes. I break possible measures up based on three categories: Expected vs. Actualized, Internal vs. External, and Counterfactual vs. Noncounterfactual. I will go through these but mention upfront that I believe the measure to associate with both Effective Processes and Effective Information are Actualized and External, where the former is noncounterfactual and the latter is counterfactual.

Before describing these different categories, I use an example to provide some intuition on how Effectiveness differs from Shannon and Semantic Processes. While it is unclear whether Semantic and Shannon Information is greater or not for observing a nearby shark vs. a dolphin in the ocean, often the observation of a shark would provide more Effective Information for the agent. Detecting a shark can enable more effective future behaviors, leading to more desired states (e.g., uninjured states). Our own semantic processes are heavily constrained by evolution and learning and, in turn, our information processing capabilities. What patterns are detected in the environment/messages⁵³⁸ largely depend on their past effectiveness for coping with the world to achieve survival states and ultimately states associated with passing on our genes. These constraints affect everything from the colors we perceive and other object properties under different conditions, to the "colors" of our emotions, to even how we reason. Thus, understanding how to measure and optimize Effectiveness is necessary to constrain how systems ground messages and process information. If we want AI to model the world similar to humans, then we will need to understand these constraints and enable them for AI by providing appropriate environments and feedback.⁵³⁹

Expected vs. Actual: Since Effectiveness constrains the nature of Semantic Processes for an agent, it should be based on the *actual* effectiveness that occurred given messages. An agent or external modeler can always use this information to learn to estimate how effective future messages will be, just as they can estimate the expected Shannon Information for outcomes generated by some modeled process. However, the value of these estimates only matters for

⁵³⁸ Although it is not entirely accurate, I will tend to use 'environment', 'state', and 'message' interchangeably in this section, depending on the emphasis.

⁵³⁹ *Research: If humans are biased towards detecting certain patterns over others, determine a way for humans to label what they see as regular and irregular in random and not so random images. The set of properties in the detected patterns can then be weighted higher for machine learning algorithms. The idea is to get AI to be biased towards the basic patterns that humans are biased towards detecting (if any). This could be crowdsourced since it requires no expertise.*

actual effectiveness to the extent that they change the actual effectiveness resulting from the messages.

Internal vs. External: An agent can intrinsically assess how effective messages received were for them. However, I claim such an assessment is fully covered by Semantic Processes, or Shannon-like Information when explicitly assessed. An external Effectiveness measure is conceptual, but it can capture how messages within the environment of the agent affect the agent's ability to achieve certain environment states. An agent does not need to itself conceptualize and measure this effectiveness to be affected by what it measures. If the agent adapts (or evolves over a lineage), then this impact of messages can constrain the Semantic Processes that emerge for the agent,⁵⁴⁰ including the perceived value or significance grounded with certain message elements. This is how Grounding emerges in line with functional purpose. It should increase the potential to meet some objective(s). Of course, agents, including humans, can not only ground messages in a perceived effectiveness quality (e.g., desire, satisfaction, fear, helplessness) but even explicitly reason about them (e.g., weighing pros and cons of a taking a job given a message about the company's latest stock price).

Thus, one should expect the effectiveness that messages have on agents to correlate with how effective the agent believes they will be, such that they will tend to act appropriately given those messages. The difference can be understood by considering that Darwinian Evolution adapted humans to ground symbols in how valuable they believe certain stimuli will be. Receiving a message that a watering hole is nearby can elicit a symbol with grounding that includes a quality of desirability or utility for a message itself. This is Effective but not the same thing as the actual Effectiveness the message really is for the agent. Note that, even though Effectiveness is an external measure, it still measures how effective a message is for the agent of interest. I claim that **experiences of effectiveness are Semantic properties, while what we want to measure for Effectiveness is not itself experienced.** To attribute what is experienced to semantics and effectiveness would be to overdetermine the experience. Instead, effectiveness is an external constraint on what is experienced. This is consistent with an understanding of how our own values are products of some larger process, such as evolution and/or culture. In a sense, there is value in our values. Similarly, an AI agent could learn to form certain values separate from the value function that drives it, as long as those learned/evolved intrinsic values increase the agent's external effectiveness.⁵⁴¹ Effectiveness is not something the agent must process themselves but instead is measured as a constraint on the agent. A message can be more or less valuable if it changes future states that the agent achieves. Thus, a key difference from Semantic Processes is that Effectiveness, while still dependent on the receiver, is not fundamentally measured by the receiver itself.

Counterfactual vs. Noncounterfactual: Effectiveness as an Actual and External measure can be either noncounterfactual (e.g., an absolute reward that a message has for the agent) or counterfactual (the reward relative to not receiving the message or relative to other possible

⁵⁴⁰ Assuming the agent meets the criteria to possess Semantic Processes.

⁵⁴¹ Or at least do not significantly decrease the effectiveness.

messages). A noncounterfactual measure, such as a thresholded fitness function to determine if an agent survives, does not require explicitly considering all of the other things the agent could have done or what policy would have let them survive. I will call noncounterfactual measures **Effective Processes** and counterfactual measures **Effective Information**. When describing both, I will continue to refer to the ‘Effectiveness’ of messages for the agent. We will see that Effective Information seems to be a generally more useful measure.

Like Shannon-like Information, as an external measure, Effectiveness can be defined by whatever we decide we want to measure as external observers. However, what objectives we choose absolutely matter for the agent’s development if the agent is learning based on the effectiveness achieved. If we want something closer to our own effective processes, then we need something that can capture the same constraints.⁵⁴²

This ability to choose what objective to associate with “effectiveness” means that effectiveness is more general than one might assume. I believe the best way to think of it is as defining X in the statement “In order to X ”. For example, if X is achieving a job promotion, then messages that help an agent achieve the promotion will have positive effectiveness. If X is to achieve higher accuracy and prediction for a supervised learning model, then images that improve test accuracy or precision and recall can be considered effective. If one only cares about training performance, then the loss function used in training can be associated with Effectiveness. Thus, it is important to not get hung up on the names we assign to objective functions. There is nothing intrinsically more effective about “reward”, “utility”, and “fitness” functions than a “loss” or any other function that can be used to compare distance from some defined objective.

Depending on defined objectives, many times the influence a message has on achieving the objective is not readily apparent. Many other messages could occur in between the time when the message is received and the impact that message has. For example, an agent may have bought a tie for a wedding that they then had available for work that contributed to receiving the job promotion. This creates credit assignment⁵⁴³ problems for assessing each message against a given objective. For this reason, it is practical to use Effective Information as a counterfactual measure, where one can estimate what would have occurred were the message not received in the same way and everything else was the same.^{544,545} In this manner, a message could have positive effectiveness even if the sum effectiveness for an agent was negative, because the sum of the messages that occurred is negative. Note that the counterfactual is not simply for receiving the message. By “not received in the same way”, I mean to include different ways the message

⁵⁴² Options to move beyond predetermined objective functions for AI will be discussed in a future manuscript.

⁵⁴³ That is, how much credit (i.e., contribution to effectiveness) to assign to messages.

⁵⁴⁴ Measuring counterfactuals is easier said than done, as both the actions and states can change. Also, the influence of messages for humans can manifest at any time, even decades later. This is a good place to reiterate that Nature does not need to fundamentally compute counterfactuals itself, as it does not need to measure its own effectiveness.

⁵⁴⁵ Counterfactuals are not the only way to assign credit in AI, but this does not matter for points made here.

could have been grounded. However, it is unclear how to capture what is essentially an infinite counterfactual space.⁵⁴⁶

Even simplifying to the absence of a message for the counterfactual case is difficult to accurately measure. No credit assignment approach is likely to be perfect. The number of messages and environment states that must be controlled for in a counterfactual process to determine their effectiveness is often intractable. Thus, there is some uncertainty involved in the measure. For some environments and objectives, there may also be uncertainty about how much of some objective should be associated with a given state or what states occurred. These are practical issues to address.⁵⁴⁷

Although not necessary for effectiveness to be measured,⁵⁴⁸ generally an agent must be able to learn given measured effectiveness as feedback (e.g., measuring how much a message ends up reducing the agent's total error or how much it enhances its future survival given their influence on predictions or behaviors, respectively). Learning is generally beneficial when optimal actions (and for Effectiveness, how valuable messages will end up being) are unknown. While unknowns are often associated with the rewards or state transitions that will occur given actions, the very nature of Semantic Processes as subsymbolic processes makes learning essential.

But what does it mean for an agent to be constrained by the effectiveness of messages, rather than the effectiveness of actions? Reinforcement Learning (RL) and other methods often base rewards on (action, state) pairs.⁵⁴⁹ This is a topic worth elaboration given other claims in this manuscript. I have already claimed that Semantic Processes fully determine what messages symbolize and their associated grounding. Thus, how messages are “processed” by agents depends on Semantic Processes already in place for the agent.⁵⁵⁰ The key claim is that **Effectiveness of messages depends on how they are grounded via Semantic Processes.**

⁵⁴⁶ (Kolchinsky and Wolpert 2018) provide, what I believe is a potentially useful measure for Effective Information (even though they describe it as measuring Semantic Information).

⁵⁴⁷ Dealing with credit assignment is an old problem. Backpropagation is a common solution used to address this problem for Neural Networks, but backpropagation over time, while significantly improved with recent advances, can diverge from the most appropriate credit assignments over long and complex histories.

⁵⁴⁸ Non-learning agents can also achieve some objective that depends on messages received. Also, the system of interest need not meet the subsymbolic criteria claimed earlier, and therefore does not need to possess Semantic Processes. It merely must be describable as something that can be states that depend on messages received by the system.

⁵⁴⁹ In RL, and Markov Decision Processes in general, ‘Value’ is defined with respect to how much total (discounted) reward each state is expected to result in. However, the Effectiveness concept being described here is associated with what is actually achieved given a message/state, and thus more similar to what would be described as the ‘Return’ in RL / Markov Decision Processes. Effectiveness depends on agent actions but is measured only for messages themselves. The idea of measuring “effectiveness” of data itself is not new. Research on the Value of Information (Howard 1966) is directly relevant. Also note that there is a difference in a message and what an agent observes. Effectiveness could be measured in either case. Effectiveness for both can depend on how an agent grounds messages, but they are formulated differently, since the latter would be a measure of effectiveness where the grounding is part of the observation state itself (rather than, say, a hyperparameter or part of the action space).

⁵⁵⁰ Or predetermined by an external modeler, as is done for ERMs. Again, since Effectiveness is an external measure, it can be measured for anything whose changes depend on messages, regardless of whether the models can intrinsically process the message. Of course, achievements would be very limited without Semantic Processes.

Therefore, Effectiveness of actions given states both drives and in turn depends on how effective messages will be. Just like Semantic Processes, the Effectiveness of messages depends on the agent, including their particular Semantic Processes that process the messages. For example, given the exact same message, an agent with Semantic Processes that cannot detect faces might perform worse than one that adapted to detect faces, as long as the detection of faces led to better states. The effectiveness of the same message could easily differ for both cases. It is therefore important to modify Semantic Processes based on effectiveness.

Effectiveness of messages is a means to an end for the agent.⁵⁵¹ For many objectives, agents that maximize the effectiveness achieved from messages will also score high on effectiveness of *(state, action)* pairs.⁵⁵² Regardless, it is the former that is closest in line with what I believe Weaver had in mind. By associating Effectiveness with impacts on objectives, the associated message must have some causal power on what the agent does (i.e., their action), which in turn has causal power on the environment (i.e., state) that impacts the agent's achievement of the given objective.

Either way, an agent must have the capacity for Semantic Processes to be altered based on their impact to agent effectiveness given state, action policies. The agent's current Semantic Processes should influence behaviors, which in turn affect the states achieved, so that Semantic Processes are modified and reassessed based on the effects those modifications have. This necessitates a role for effectiveness of messages in learning. Messages for a non-learning system may have effectiveness, but generally the agent/system must be able to learn based on the effectiveness of messages for it to have any significant impact, particularly on the utility of the agent's Semantic Processes.⁵⁵³

⁵⁵¹ I would imagine that ultimately agents must learn in a way that modifies the embeddings (e.g., underlying neural network properties) that ground messages they receive from the environment, and where the embeddings influence the policies (i.e., determine the perceived state of the environment). Evolutionary processes seem appropriate for adapting innate grounding properties (and/or innate propensities during learning) in the same manner. Whatever this process is, the approach should not involve attempting to modify the grounding of each individual vector but instead adapt more general constraints on the underlying semantic space. The grounding of some objects may be more influential on this evolution than others, but, in general, the symbols we generate are not each individually adapted through evolution.

⁵⁵² *Research: It would be interesting to develop a reward function for somehow finding the most patterns/regularities in environments. This may simply be the expected information gain, but how "pattern" is quantified should include predictive power and might be proportionally more rewarding for longer time intervals. A benefit is that this can be done with a robot that explores by interacting with the real world and its objects, constantly seeking out more information. To maintain curiosity, patterns already discovered could decay in reward, and exploration should be strong enough at times, possibly during "youth", to warrant "perturbing" known patterns to see what else can be derived. Also, actions available should significantly impact what patterns are capable of being actualized. It would be interesting to see if a tradeoff occurs at some point (one can add an energy cost to degrees of freedom in a virtual robot that can genetically mutate new motor options, and maybe for sensing too). Also, consider how knowledge in the form of discovered patterns can ratchet up discovery of otherwise unlearnable patterns. This would emulate how knowledge builds on previous knowledge.*

⁵⁵³ *Research: Consider how Abstraction relates to the Law of Requisite Variety (Ashby 1958). Abstraction should reduce variety, but the cost needs to be considered based on Effectiveness, including across evolutionary history.*

By coupling the learning of Semantic Processes to how much they increase agent effectiveness, Semantic Processes can in theory become more “human-like”, as needed for AGI. In turn, those processes can play a more useful role in constraining agent behaviors. As far as I know, this type of coupling has not been considered, including in use of embedding methods for Natural Language Processing (NLP) research. Despite the breakthroughs in NLP using embedding methods (e.g., (Devlin, et al. 2018)), it is unclear how much of the Effectiveness they capture in this process of passively ingesting text data. A similar case can be made for any Deep Learning methods that are passively trained as discussed in Chapter 2. Errors made by Deep ANN models are often nothing like errors humans would make (e.g., classifying people escaping a flood as a group of people standing on the beach (Lake, et al. 2017)). It is plausible that a major reason for this is that Semantics alone does not sufficiently account for the significance of the different potential features and how they are most effectively perceived. Significance for biological intelligence starts⁵⁵⁴ with evolutionary impacts.⁵⁵⁵

That said, I recognize that incorporating evolutionary adaptations to achieve semantics similar to humans is a hard problem. We do not have that history available, and, even if we did, it would be impossible describe conceptually (i.e., explicitly) and too complex to fully model. It is highly improbable that some simulation environment could be expected to guide AI agents to evolve as we did. Clearly, to achieve constraints that align with those of humans, one must approximate or find alternative approaches. There is surely existing research that can help understand how constraints emerge for agents. For example, Donald Hoffman’s lab has tested their “Fitness Beats Truth” (FBT) theory in simulations (Prakash, et al. 2020). The research demonstrates the relevance for what I am generically calling “Effectiveness”. Hoffman demonstrated in evolutionary models that agents guided by fitness and not accuracy with respect

⁵⁵⁴ Actually, I plan to discuss even more fundamental processes in a separate manuscript.

⁵⁵⁵ It is possible that at least some remnants of effective-driven biases are inherent in human-derived data. For example, the words used by humans may indirectly reflect evolutionary biases of the humans who generated them. However, this is likely a very weak signal, as there is a difference between what is said vs. how what is said is grounded.

to their reality survive, whereas agents with more accurate representational constraints go extinct.^{556,557,558,559,560,561}

The main point I want to make in this section is that Effectiveness is as fundamental as Shannon and Semantic Processes. AGI will not be born from passive observations or surface level modeling of psychological and cognitive theories as is done with typical Cognitive Architectures. If Hoffman can go as far as hypothesizing that FBT warrants consideration of Idealism (Hoffman 2019), then surely the implications for Effective Processes' role for AGI are deep.^{562,563} Research combining Embedding methods with Reinforcement Learning, such that agents' decision making

⁵⁵⁶ It should be noncontroversial that agent evolution should only "care" about accurately perceiving reality to the extent that doing so improves fitness. A challenge in our own world is how we can actually assess the relative importance of fitness and truth for ourselves, given our own fitness-colored perceptual and conceptual adaptations.

⁵⁵⁷ **Research: Determine what would be required for Hoffman's models to develop Semantic Processes as shaped by fitness.**

⁵⁵⁸ **Research:** Given Hoffman's FBT theory and possible implications for denying the truth of its own claims, consider designing and developing a method that uses fitness to evolve/adapt perceptions that recursively converge over themselves. This may be an interesting means to create concepts and then reasoning, since this may be exactly how and why conceptual processes evolved (i.e., to create and improve higher level assumptions about a "truth" given lower-level perceptions). For example, consider a basic FBT model that creates perceptions, and then consider the conclusions of that model, and how those conclusions, as the next level of perceptions (concepts / meta-perceptions), would be incorporated into the model to be shaped by fitness, and so on. What would this all converge to?

⁵⁵⁹ **Research:** Set up experiments where the FBT theory is evolved for agents for perception just like Hoffmann's research showed. However, evolve it to where one can test a requirement that the agents also perform the same experiment themselves, though in a simpler way. Specifically, the agents must learn that their perceptions do not mirror reality. This would be an interesting way of testing FBT and has philosophical implications.

⁵⁶⁰ **Research:** Consider if there are any implications from FBT for information itself. "Error" assumes a correct signal (i.e., some truth), as does Shannon coding and, really any type of prediction model. Can we measure whether error correction in the brain, and associated models of cognition (Friston 2012) corrects as would be expected under the assumption of optimizing accuracy? Instead, is there a correction factor that must account for evolutionary fitness? Or perhaps the better description is that fitness maximization is the best fitting model with some correction factor for a general bias towards accuracy? These questions hint at the need for formalizations for Effective Processes that distinguish it from Shannon Information. Consider what it means to compress and "error correct" in ways that are effective with respect to fitness rather than an accurate representation of the world.

⁵⁶¹ **Research:** (Laughlin 1981) *measured Shannon Information of the fly eye based on contrasts in natural environments. Can the same be done for Effective Information in a model? That is, can we associate stimuli with how important an encoding of it is to the organism's fitness? Consider what this would be. Then, if the model evolves Semantic Processes, it can be measured in an effectively grounded context.*

⁵⁶² **Research:** Consider if there is any value in trying to infer/predict from observational data of an evolutionary model. For example, this could be a classifier or embedding method processing pixels from an evolutionary algorithm to determine the lower dimensional manifold upon which evolution occurs. The evolutionary process can be restarted an arbitrary number of times, and so the manifold should be quite comprehensive. The manifold can then be compared to a similar manifold learned for the genetic algorithms (i.e., the "genes") of the agents. Do they have similar/correlated dimensions? Is a universal mapping function possible from one to another? If so, can that same mapping function be used on real genetic data? This may just be an exercise in arbitrary basic research but seems interesting at least to spark ideas given the process and analysis applied.

⁵⁶³ **Research:** Consider how datasets could be created that help to reverse engineer perturbations due to Effective Information. The goal is to see if patterns specific to evolutionary history could be inferred (i.e., reverse engineering evolutionary history) so they can be used to bias pure Semantic Process learning (e.g., neural embeddings from text corpora). For example, one could create an evolutionary model and control for adaptations and then compare the data that comes from the different evolutionary branches.

is influenced by their semantic embedding space, is worth testing to see if it improves the semantics learned vs. effectiveness-independent embedding methods (if we can determine useful objective functions).⁵⁶⁴ Even if it does not get us close to human-like semantics alone, the progress should provide useful applications and insights on associated AI problems.⁵⁶⁵

It is worth noting that, even though Reinforcement Learning has demonstrated some of the power of performative intelligence by dominating games like Chess and Go (Silver, et al. 2018), these games are not ideal demonstrations of the full potential of Deep Learning to learn implicit features. Deep Learning detractors focus on the fact that these games are bounded in contrast to real world activities. However, I believe there is a deeper simplification in these demonstrations in that they are *Conceptual* games. They are completely determined by explicitly defined rules. This means that, in theory, there is some explicit model that can perform optimally. Feature Learning methods still perform better due to their ability to scale through generalizable function approximation, which includes finding more abstract implicit features. But there are, at least in theory, optimal explicit policies. Many games are completely determined explicitly, such as card games and board games. Many activities, however, typically involve real-world actions that are not constrained by explicit rules (e.g., move right one space). Explicit determination is rare for real-world problems, at least beyond the level of primitive physical mechanisms.⁵⁶⁶ There is no guarantee that a purely explicit model is theoretically optimal. My conjecture is that most real-world activities involving general intelligence are closer to these cases, and cases where purely explicit models are optimal are the exception.

To close this section, I discuss how Effectiveness can uniquely compress data (**Effective Compression**). Evolutionary and lifetime experience constrain Semantic Processes to choose specific compressions from the potentially infinite number of choices. **Effective Processes compress messages by removing effectively irrelevant elements, or more generally optimizing the loss vs. length tradeoff for impact on effectiveness.** In this way, Semantic and Effective Processes both relate to optimizing tradeoffs between compression and information loss in how existing knowledge is applied to messages.⁵⁶⁷ Effective Compression also likely influences how much information is stored in the brain for Semantic Processes vs. how much must be received via the environment.⁵⁶⁸ What is stored is stored because it facilitates how information from the environment is processed, including model formation. Thus, Effective Compression helps

⁵⁶⁴ **Research: Do the research to constrain Semantic Vector Spaces with Effective Processes by implementing them in evolutionary and/or reinforcement learning agents.**

⁵⁶⁵ *Research: Develop a method to analyze AI agent internal models and compare their communication effectiveness to the semantic vector space they are projected onto to see what information loss in the semantic space impairs using them directly for communication. This facilitates building in structure, either as explicit rules or state space constraints, that recursively increases internal Order with the representations themselves. How Order is measured must also be determined.*

⁵⁶⁶ Just imagine trying to build a purely explicit model for a robot to learn to create beautiful new paintings or be a funny comedian.

⁵⁶⁷ *Research: Relate Effectiveness to Risk, which I think can be generalized to the probability associated with how many options are expected to be available to exploit new information in the future. This is why one wants to exploit more than explore when risk is higher for exploring (and vice versa).*

⁵⁶⁸ The relation to the concept of Stigmergy will be also considered in depth in a separate manuscript.

determine not only how information is grounded, but what information is grounded. Not storing information risks that it will be unavailable when needed. The tradeoff is presumably related to focusing processing time where it is more relevant, energy involved in storage, possibly energy costs for processing or execution time, and interference with grounding of other stimuli.

Effective Compression does not seem to require explicit symbols. For example, emergent heuristics like motor reflexes and behaviors, such as evolved or learned action sequences, reduce agent exploration requirements (e.g., every possible combination of action sequences given every possible stimulus) for effective learning. **Heuristics reduce state space size, which should be associated with Effective Compression.**^{569,570,571,572} Since Effective Compression can occur for unconscious processes, such as motor reflexes, it itself cannot be directly experienced. This warrants not directly associating it with Grounding. Note that the general concept for reflexes can be translated to sensory heuristics, including mental chunking. Without understanding that Effective Processes are not measured for the grounding itself, it would be difficult to disentangle what is Effective vs. Semantic Compression. Effective Processes constrain Semantic Processes, just as they constrain physical behaviors for an agent.^{573,574}

⁵⁶⁹ **Research: Determine how to measure Effective Compression.**

⁵⁷⁰ Using known heuristics as guideposts for developing AI heuristics may help steer AI to be more human-like.

⁵⁷¹ *Research: Consider evolving/learning better compression algorithms by maximizing fitness or reward for agents by giving them random data and letting them determine how much they can compress it. This could be done systematically by requiring the algorithm to maximize compression with all sorts of data, or one could evolve/adapt different winning algorithms for different data types, data sizes, or with different correlation magnitudes. After compressed, the algorithm must decompress it to receive reward. For optimizing for lossy compression, the reward can be weighted by the compression amount and the lossiness. One could also factor in how long it takes to compress and decompress into the fitness/reward function. Then, consider exploiting Semantic Processes by enabling agents to coordinate by sharing certain knowledge, maximizing the amount of compression enabled by the minimum amount of shared data. This should learn what information to share to enable the most compression across data messages, which then for the real world would require that same information to be stored on any computer that will decode/decompress data encoded/compressed using the discovered compression algorithm.*

⁵⁷² *Research: (Wolfram 2002) demonstrates that certain cellular automata (CA) types (e.g., Rule 30) are irreducible. Consider if there is a possible experiment that would allow Semantic and Effective Processes to determine some optimal emergent level to describe what is observed in the different CA rules (e.g., a lossy compression of a rule as “irreducible” or exhibiting certain macro-objects).*

⁵⁷³ *Research: Consider AI research to develop such heuristics. Reflexes can be instantiated in Reinforcement Learning (RL) by implementing a modifiable coupling strength between actions, such that sequential or simultaneous actions that tend to work well together have additional coupling parameter (e.g., probability, force, timing) weights learned. Determine how this could then be applied to develop sensory heuristic “reflexes”, such as a coupling that develops attractors in Semantic Vector Spaces to bring objects together (as fixed points, limit cycles, etc.) based on effectiveness in the world. Determine the coupling equations to implement with RL.*

⁵⁷⁴ While I emphasize the influence that Effectiveness has on Semantic Processes, the Semantic Processes also would clearly influence behaviors (e.g., policies that define the probability of how the agent will respond to difference perceived environment states). The influence on Semantic Processes means nothing if it does not in turn influence the agent’s behaviors. There is a lot to consider for effective policies. For example, a very general constraint may be to evolve policies that tend to increase freedom (i.e., the number of choices available given future states). This assumes that agents are generally good at choosing among possible actions or states with respect to objectives. The relationship between Effectiveness-driven Semantic Processes and Policies is worth exploring for AI. For example, understanding how grounding messages with certain emotion qualities, how uncertainty is perceived, and whether

Candidate Descriptions for Concepts Related to Information

This section provides an incomplete list of incomplete descriptions for various AI-relevant concepts based on the claims made in this chapter.⁵⁷⁵

Messages: Data received from the environment, including from other agents or internally (e.g., propagated from other neurons). **Message elements** are distinct components in a message. They can be defined by the transmitter to represent individual symbols, but different receivers may group them differently as symbols.

Microstate: A set of variables and their values upon which a message can inform. ‘Microstate’ is used as theoretical device in this manuscript to describe what may be an implicit construct.

Macrostate: A set or distribution of possible microstates. I often use ‘**model**’ interchangeably with macrostate in this chapter. ‘Macrostate’ is used as theoretical device in this manuscript to describe what may be an implicit construct.

Reference Frame: The perspective of a theoretical receiver for which messages are perceived. This is also related to what would typically be associated with different emergence **levels** or scales for concepts.

Shannon-like Information: Information is a vague, or at least debated concept. I instead refer to ‘Shannon-like Information’, which is a reduction in uncertainty a message provides a receiver, measured by the change in a macrostate with respect to a microstate. Information is symbolic and conceptual, since it is always about something (i.e., the microstate). Information requires an observer and cannot be described as possessed by a system if the system does not possess Semantic Processes. However, it can be measured for any system by an external observer. ‘Shannon-like Information’ covers Shannon Information, Algorithmic Information, “Semantic” Information, and any other theories that meet the definition provided here. Shannon vs. different Semantic Information Theories can be loosely distinguished by their types of referents. What they share is that processing is measured over the symbols within their given reference frame. I would prefer to propose that my claims for Shannon-like Information cover Information in general. However, I decided that debating whether the claims cover all the folk uses of the term would distract from the points in this chapter.

Semantic Processes: Semantic Processes form the micro and macrostates used to measure Shannon-like Information and is synonymous with (Symbol) Grounding. This can also be thought of as constructing the indices for Shannon-like Information processing of messages. It is subsymbolic, meaning it is not described by the indices it constructs for the Shannon macrostate or any other explicit indices for its own reference frame. While it is a measure for the message, it is receiver-dependent. Semantics is always projected from a receiver’s Knowledge System to the message, and what is in each Knowledge System will change the grounding and therefore

the grounding of messages should influence whether a given agent chooses to explore or exploit policies, should all affect behaviors. Indeed, how an agent forms micro and macrostates to assess their concept of effectiveness of possible messages and decisions can only be accounted for with Grounding but should be constrained by the external effectiveness to be of any utility to the agent. Hopefully, these examples demonstrate the potential for accounting for Effectiveness and Semantics for future AI.

⁵⁷⁵ **Research: Mathematically or algorithmically formalize each of the concepts described in this section.**

information (i.e., the same message can have different information quantities for receivers with different existing knowledge).

Semantic Specificity: A proposed magnitude of semantics for a message. It can be measured in a Conceptual or Semantic Vector Space as the Specificity of the processes evoked by a message and is a combination of the **Complexity** (number of quality dimensions involved) and **Quality-Level Specificity** (average range or Shannon-like Entropy along each quality dimension involved). Together, these are measured as the “activated” hypervolume within the receiver’s vector space. Greater Semantic Specificity is associated with more quality dimensions and smaller quality ranges (or Shannon-like Entropy).

Semantic Information: Information from the reference frame where messages are symbols associated with meanings, where ‘meaning’ is typically associated with some state of affairs as opposed to some basic set of distinct forms. However, since both require first understanding something as symbols, Semantic Information Theories are no more related to Semantic Processes than Shannon Information. I therefore group them into what I call, ‘Shannon-like Information’.

Effective Processes: A measure of the absolute impact of a message to some given objective for a receiver. By necessitating some impact to the agent’s behaviors, effective processes have causal power on the agent and the environment via the agent.

Effective Information: A measure of the impact of a message to some given objective relative to counterfactual cases where the message had been grounded differently by the receiver (including not grounded at all). By necessitating some impact to the agent’s behaviors, effective information has causal power on the agent and the environment via the agent.

Knowledge System: Knowledge is assumed in this chapter to refer to the totality of what an agent has available to use to ground messages. ‘Knowledge System’ could be described as both this content and the Semantic Processes applied to it, including to form explicit concepts. Knowledge is fundamentally implicit and relates to the ability and propensity of a receiver to form different micro and macrostates given messages. Thus, the Knowledge System is not merely a model of symbols and their explicit relationships, in contrast to ERM Knowledge Models.⁵⁷⁶ Forming the micro and macrostates can generate explicit concepts. The set of all potential explicit concepts that can be formed could be described at the Explicit Knowledge of the agent. Intuitively, Knowledge enables comprehension of messages.⁵⁷⁷ This description does not limit knowledge to be true, which would be more consistent with the intuitive concept of Knowledge. One option is to differentiate Knowledge as what is truthful from another measure for what is described here (e.g., Knowledge Ground). I leave this for further consideration. There is in theory

⁵⁷⁶ Note that I use Knowledge ‘System’ to avoid conflating the term ‘Model’ with the model at the level of a given macrostate. A Knowledge System could be thought of as a meta-model or model space, as long as this is not conflated with the assumption that all possible models are explicitly about what they model from the reference frame of the Knowledge System. Knowledge is assumed to be fundamentally implicit, as is required for Semantic Processes to use the receiver’s knowledge to form micro and macrostates.

⁵⁷⁷ *Research: How an agent’s amount of Knowledge is measured must be determined. Should it be selecting the “best” or “most appropriate” macrostate for a message? Should it be the Specificity of the hypervolume given the message, and therefore the same as Grounding? Should it be how much Shannon-like Information the knowledge was able to actualize for messages?*

a semantic measure of an agent's total knowledge, as well as the knowledge gained by a given message received. I leave it to future work to determine how to measure these properly to account for its implicit properties.⁵⁷⁸ Measuring just the explicit knowledge may be equally difficult, since one must account for the generative nature of knowledge itself, which is generated from the subsymbolic processes.⁵⁷⁹ Of course, this issue is no different for attempting to measure any human's knowledge.

Thinking: **Explicit Thinking** is a path dependent process for an agent that iteratively encodes and decodes internal messages. By generating messages, Explicit Thinking is experienced as talking within one's own head in natural language. **Implicit Thinking** does not encode associated Semantic Processes into explicit messages with associated symbols. It remains subsymbolic. Both forms of thinking involve implicit processes and are thus associated with experiences. Specifically, Implicit Thinking can generate Explicit thoughts. Explicit Thinking as forming iterative macrostates is thus functionally distinct from directly manipulating symbols to form symbol chains. Thinking, as described here, is a prerequisite for a system to be considered a Cognitive System. One might place more requirements, such as thinking involving useful inferences, but I believe such cases would refer instead to Problem Solving. Regardless, further requirements do not conflict with the claim that Thinking, and therefore Cognition, at least require Semantic Processes.^{580,581,582}

⁵⁷⁸ *Research: Determine how these would be measured. Once a way to measure it is determined, can it, even in principle, be summed over all possible messages or activations to measure the total knowledge of an agent? A related measure might be something like the Total Semantic Specificity or something that captures the number of vector space positions or geometries that differentiate between possible messages.*

⁵⁷⁹ *Research: To constrain implicit knowledge to truthful beliefs, consider if it could be approximated by conditioning the explicit assertions the system generates to those that are truthful, and then use the percentage to weight the total measure of implicit knowledge (assuming a statistically valid sampling process can be performed).*

⁵⁸⁰ *Research: Identify if there are conditions that determine whether grounding measured over conditional sequences of data (e.g., sentences), are less or greater than the sum of the grounding measured for each data element (e.g., each word).*

⁵⁸¹ As mentioned, (Haken and Portugali 2014) relate Abstraction and other processes to what they describe as Information Inflation and Deflation, as well as adaption between Information Inflation and Deflation. This form of adaption is something we likely do all the time, often iterating between Inflation and Deflation. If so, then abstraction is more frequent than what we would typically associate with Reasoning and perhaps Thinking in general. Although I do not describe these processes further, I find them compelling and refer the reader to the associated reference.

⁵⁸² Although not high on the list of human processes to model in AI, generating Streams of Thought seems like an expensive process if it has no function. There is a conditional aspect of Grounding in that it "makes available" vectors or vectors subspaces to condition downstream thoughts. Perhaps the relations that we define as grounding may be variational processes occurring stochastically due to, and as part of, the process of making a vector available. If activation of a neuron or neuronal circuit were associated with a vector or vector subspace, then it may have stochastic activations, the tendencies of which make up their grounding. For example, if I receive the message 'dog', the associated vector may have stochastic processes that lead to different vector changes that may have different conditional probabilities, such as evoking a visual grounding for the Disney character, Pluto. Activations would therefore involve vector changes probabilistically conditioned on messages. In some mental states, the conditional probability distributions might be globally made more uniform, leading to more random thought paths. This is just a guess for how streams of thought occur.

Reasoning: I provide only a basic description of Reasoning here.⁵⁸³ Reasoning involves Thinking constrained by a goal⁵⁸⁴ state(s), such as satisfaction criteria. As for thinking, any explicit processes must be encoded and decoded by the subsymbolic processes by forming micro and macrostates. While it is natural to think of Semantic Processes as generating explicit micro and macrostates,⁵⁸⁵ this is not necessarily their primary function. Consider that, when we communicate with others, we often must seek common ground. In contrast, when communicating to ourselves internally, we seem to always understand the meaning of the explicit concepts we generate. If anything, we struggle with the opposite, which is to find the “right words” to express ourselves. Perhaps this demonstrates that explicit properties, while common, are not primary for reasoning itself. They can perturb the grounding, perhaps even helping find certain implicit thoughts, but they always ride on top of it, where the primary “inferencing” occurs.^{586,587,588} Use of Semantic Processes enables the ability to form higher-order symbols, grounded in what is evoked by and generates other symbols. Therefore, abstract concepts for higher-order semantic reasoning are possible.⁵⁸⁹ Using a Semantic Vector example for Reasoning to answer a question, the receiver performs operations over beliefs to arrive at a result vector determined to meet the goal state conditions. Reasoning can reduce possible interpretations

⁵⁸³ A larger discussion on Reasoning is planned for another manuscript.

⁵⁸⁴ “Goal” is used very generally here. Simply put, it is satisfied by any implicit or explicit state that serve as stopping criteria for the associated Reasoning process.

⁵⁸⁵ I have unintentionally contributed to this assumption by emphasizing this role for Semantic Processes throughout this chapter. It served as a crutch to explain Semantic Processes given the need to explain it with symbols.

⁵⁸⁶ I put ‘inferencing’ in quotes because it is unclear that this is the appropriate term at a subsymbolic level. However, I cannot think of a better term for processes that lead to states that satisfy certain properties that could be extrinsically perceived as inference.

⁵⁸⁷ However, Conceptualization does seem critical for higher-order thoughts. Thus, the balance between purely subsymbolic and symbolic inferencing is unclear.

⁵⁸⁸ *Research: If explicit concepts capture our narrative self, then consider if there are implications for the famous Libet et al freewill study (Libet, et al. 1983). This study found that subjects report making conscious decisions about a half second after their measured brain activity reveals the decision. Could the early brain activity be associated with a conscious but implicit semantic grounding (i.e., a percept associated with the act of deciding), while the time humans report they decided is associated with the forming of explicit concepts (i.e., self-narration) of that grounding? In other words, there is a delay between the grounding and the conceptual transformation involved in being able to report what one perceived or did (even to oneself). It is plausible that we have adapted to perceive perceptions and the conceptual form about them as simultaneous for better mental coordination (e.g., not recalling experiencing the same grounding twice, once for encoding and another for a decoding evoked during explicit self-narration). This would not mean that subsymbolic processes are nonconscious, only that we miscalculate their timing. More generally, Semantic Processes as percepts would suggest any experience of a decision might always occur implicitly. Still, if such coordination was valuable, it would seem even more valuable to simply delay the subsymbolic experience, rather than forget that we already experienced it. If so, then we would not be conscious of it at the time it occurred, validating the initial conclusion of Libet et al’s study. Then again, it would be useful to experience it earlier if it helps react quicker when needed. This functional difference could be possible to control for in experiments. It would be interesting to implement experiments that could assess whether there is any delay between grounding and conceptualization, in what circumstances, and whether both can ever be separately reported as experienced.*

⁵⁸⁹ This sort of reasoning chain includes interpretation processes of complicated messages. For such cases, the words might need to be interpreted as individual concepts first, then certain phrases, and so on to more abstract semantics. Of course, it is much messier than following this clean chain of abstractions.

(i.e., answers) given a message (i.e., question) by inferring over existing knowledge. When this process forms explicit processes, Logic is explicitly applied,⁵⁹⁰ and the associated reduction can be measured using Shannon-like Information.

Learning a New Meaning: This involves learning new grounding for symbols given a message. Decoding a message can adjust positions of vector regions in Conceptual Space. For a Semantic Vector Space, this involves modifying neural weights as is done during training with neural embedding methods. The modifications might involve moving prototype positions, adding new boundaries or adjusting existing ones, and/or adding quality dimensions.⁵⁹¹ There is plenty more to consider about what is involved in learning meanings in light of claims made in the literature and about Semantic Processes in this chapter.^{592,593,594}

Learning a New Fact: This involves activations of existing vectors with vector operations, and possibly a way to associate them with a property for implicit degree of belief.^{595,596} While current Semantic Vector Spaces are often used for single activations (e.g., return a word), it is possible

⁵⁹⁰ This does not entail that the Logic applied is applied well by the agent.

⁵⁹¹ *Research: It is unclear how new quality dimensions can be added within standard neural networks if dimensions map one-to-one with the neurons in the hidden layer used for the embedding. It would need to be possible to add neurons during training. This would also be an issue for biological neuronal networks if new neurons no longer form in the associated brain region(s). However, I am skeptical that each quality dimension maps to a single neuron in biological brains. It is worth trying to determine more biologically plausible neural representations for quality dimensions, which might also lead to improved AI embeddings.*

⁵⁹² *Research: More study, including introspective reports while learning meanings for symbols and controlled experiments for the minimal requirements for learning them, would be valuable. Existing studies and theories should be translated to Conceptual Space processes where not already done.*

⁵⁹³ *Research: As (Gärdenfors 2014) explains, some concepts are easier to learn (i.e., with fewer samples), and Conceptual Spaces can explain why. It is easier for a child to understand the new concept of chartreuse than the concept of escrow because the child has good coverage of concepts in the domain of colors and, critically, they can identify the position of chartreuse in the geometry of the color domain, whereas the domain for where escrow would be positioned has not matured in the child. This has implications for how humans can learn new concepts with few samples compared to current Machine Learning models. Figure out how to implement this for fast learning in Semantic Vector Space models to improve learning with sparse data. Classifying new objects, etc. is easier if the classification process includes grounding observations with adequate dimensions and observations to help determine where to locate new observations. This is less about training and more about inquiries and iterative mentation to achieve grounded comprehension. This could potentially include some process of energy minimization (which might be formalized in terms of frustration in continuous, high-dimensional semantic vector space, along with interactions with the environment via different physical, sensory, and communicative actions).*

⁵⁹⁴ Learning a new meaning could reduce possible interpretations for messages evoking the symbol(s) now associated with the new meaning (depending on the context it is found in). It would impact future model formation for such messages, including how new facts are interpreted (i.e., Shannon-like macrostates about some world domain), thus relating to how new facts are learned. One might expect symbols that are more grounded to reduce possible interpretations for other macrostates on average. For example, if only mammals are known, then learning about birds can reduce the possible interpretations of future messages to the smaller subset for which the messages actually apply.

⁵⁹⁵ Of course, an explicit degree of belief is also possible.

⁵⁹⁶ For theories that omit any assumption of truth criteria for Semantic Information, using the term “fact” may seem misleading. I assume here a more general case that requires some degree of belief by the receiver. As already mentioned, accounting for belief avoids the so-called “Scandal of Deduction” raised by (Hintikka and Suppes 1970). Semantic Information is gained for deductively obtained facts when the receiver’s Conceptual Space reaches conclusions that have never been activated with respect to the associated assertion.

that any message (e.g., “the soup is hot.”), especially ones not repeated sufficiently to form a single mental chunk, involve activating a vector path. A message could directly evoke the path activated (e.g., someone telling you, “the soup is hot”) or it could involve any degree of reasoning to infer facts (e.g., “if the soup is hot, then I cannot eat it yet, or I will burn my tongue like last time”).⁵⁹⁷ Activated paths need not involve learning a concept that permanently changes the underlying grounding. Learning a new fact can simply involve activating a vector or vector path that has not been activated (or was forgotten) within the continuous metric space of one’s Knowledge System. However, since biological semantic spaces are always able to learn to some degree, purely internal dynamics in the vector space have the potential to modify grounding.^{598,599,600} Once a new fact is learned, retaining it must involve some change in grounding, though it may be quite local. This manifests at the very least as a propensity to re-activate the fact given certain (internal and/or external) messages, for as long as the fact is retained in the receiver’s Knowledge System (i.e., in their memory).^{601,602}

Meta-Learning: Grounding more meanings and facts in a Knowledge System should change how well it is able to assimilate future meanings and facts. Ability to assimilate new knowledge naturally relates to Intelligence and may be enhanced with meta-learning strategies (i.e., learning how to optimally learn). Meta-learning facilitates, among other things, context switching which the brain does frequently (e.g., over seconds based on subjective self-report). Meta-learning is a critical capability for AGI. Although evolution endowed us with the ability to meta-learn, there are also processes we learn during our lifetimes that improve learning. Another key factor with human learning is the ability to make use of existing knowledge to learn more advanced concepts. Education levels fulfill such a role, which has inspired Curriculum Learning for AI (Hacohen and Weinshall 2019). Humans may be the only obvious inspiration from biology in this regard. At the

⁵⁹⁷ *Research: Consider whether vector paths themselves should be associated with certain meanings. This would be more in line with a Process-centric Science.*

⁵⁹⁸ Activating vectors that have not been activated before should be associated with a gain in certain measures for Semantic Information. This could model how new insights are generated during self-reflection.

⁵⁹⁹ *Research: These internal dynamics can be pictured as the theoretically infinite possible vector paths creating proportional continuous differences in magnitude of neural network activations or perhaps continuous changes in firing rates. Further study is required to determine where the analogous continuous properties are found.*

⁶⁰⁰ *Research: Gaining new insights need not directly involve external inputs but there is no major reason to see this as different in kind from internally generating insights; internal inputs must be able to evoke learning, including synaptic changes. The question is how to best model this in Conceptual / Semantic Vector Spaces. It is insufficient to simply search the space. Search must thus also involve the space itself changing.*

⁶⁰¹ *Research: Studies are required to determine how facts associated with working vs. short and long-term memory, as well as episodic vs. declarative memory relate to changes in an underlying Conceptual Space and therefore grounding.*

⁶⁰² In truth, it is unclear that learning new meanings and learning new facts are distinct processes. The same fundamental process seems involved in both.

level of multiple agents, evolution can be applied as a meta-learning strategy to evolve agents that learn more effectively.^{603,604,605,606}

Intelligence: Intelligence involves Reasoning efficiency and capacity. Since Reasoning relies on the subsymbolic grounding, Intelligence also involves implicit processes, although often with observable products. Efficiency and Capacity include the ability to reduce uncertainty from available information and experience. For example, for a Twenty Questions game, it is intelligent to consider whether the person with the answer is less likely to choose certain things as the answer because of their age or inferred knowledge. Such considerations can occur through active reasoning or as heuristics learned that manifest as intelligent behaviors. **There is a strong relationship between Semantics, Intelligence, and Meta-Learning since new data becomes easier to assimilate when it can be effectively interpreted. Intelligence depends on Reasoning capacity and knowledge to form a larger set of potential models to ground messages (e.g., understanding terms in Jeopardy! clues). Efficiency involves the ability to form these models as needed (e.g., quickly and reliably finding the micro and macrostates that provide effective Jeopardy! responses).**⁶⁰⁷ **Together, these enable using grounding to solve complex problems (e.g., propagation across a chain of macrostates to arrive at a solution that satisfies some criteria, such as a problem provided by a teacher or oneself).**⁶⁰⁸ **The ability to select the “best” micro and macrostates given what is available would also both maximize comprehension of messages as complexity grows and accelerate learning since more appropriate macrostates are updated.**⁶⁰⁹ **I also believe Intelligence also relates to Capacity in the sense of deeper**

⁶⁰³ *Research: Consider AI studies to measure how comprehension enables learning new facts and concepts.*

⁶⁰⁴ *Research: Consider a Hierarchy of Active Learning (Ren, et al. 2020) models to learn optimal thinking about thinking about thinking, etc. for transfer learning. A hierarchy, where each upper layer learns the optimal conditional probability distribution for sampling the lower layer, should enable some transfer learning. Doing this for active learning is just an example. Higher layers could propagate to lower levels to select the best active learning model to use for querying data. This way, as long as the highest-level active learner can query to get labels at its own level, it can select the best lower-level Active Learning model for the given use case.*

⁶⁰⁵ *Research: Use Active Learning to choose Reinforcement Learning (RL) Actions with most uncertainty. This augments the exploration-exploitation tradeoff to not only explore, but to explore in an informed way that exponentially reduces exploration actions to learn the full environment within a chosen error. This is natural to add to a Deep Learning RL methods to improve function approximation of the state space.*

⁶⁰⁶ *Research: Consider semantic vector space trajectories as a means for semantically-driven sampling for inference.*

⁶⁰⁷ It is not clear that Intelligence is limited to efficiency only with respect to Reasoning. Intuition could be said to require efficiently (or with greater probability) achieving answers or insights without explicit processes leading to them. Thus, this description for Intelligence is likely incomplete (if there even is such a thing as a complete and acceptable definition for Intelligence).

⁶⁰⁸ *Research: Consider if the best representation of Intelligence is not a model of some thing, but instead a model of change (or a meta-model for how models change). Biological intelligence seems like it may be best represented as a modeling process. Intelligence should relate to the ability adapt over time for nonstationary environments without needing to optimize to some equilibrium. If optimization must be involved, then the optimization function should be able to change over time.*

⁶⁰⁹ *Research: Create a hierarchical semantic vector space to see if the search process (speed to getting answers) is improved (which also improves communication speed) and if it is more effective in some scenarios given the lossiness tradeoff. This approach builds concepts from primitives that form new concepts (“chunks”) by putting primitives together, as well subparts, all the way through higher levels emergent concepts. What lower-level concepts to chunk might be driven by evolution (Effective Information) and learning (perhaps based on some sort of significance*

Comprehension. This includes nested concepts (e.g., he knows that I know that he knows, etc., or the depth of a series of logical operations within a given assertion, such as, it is not the case that it is not the case that p or not q is equivalent to not r). Computers excel at such logical operations but without grounding them. Humans lose the grounding for such nested concepts quite rapidly. It is possible that AI could be developed that is able to ground concepts with considerably more depth than humans.⁶¹⁰

Transfer Learning: Transfer Learning facilitates learning on new tasks or domains by leveraging knowledge gained from different tasks and domains.⁶¹¹ This can include learning with fewer samples, as well as the capability of learning concepts or behaviors that would be impossible without already understanding other concepts or behaviors. This is a major AI research area as it can help resolve the immense training requirements of Deep Learning methods. Semantic Processes, by leveraging existing knowledge to form micro and macrostates, may in theory be able to generalize beyond what is possible with Shannon-like Information by grounding across different data sources. Analogy Making is critical to Transfer Learning since existing understanding and relations between grounded concepts are used to ground new meanings and facts. Analogy Making is a core Semantic Process. This will also be important for constructing AI that can rapidly adapt from the training received in simulated environments for their use in the real-world.⁶¹² More generally, if Semantic Processes can broaden the training distribution beyond specific micro and macrostates, then the agent's span of knowledge can be applied to a significantly wider set of problems encountered. Note of course that Effective Processes are also important for Transfer Learning by constraining Semantics Processes.

Common Sense: Common Sense is the subset of knowledge, and how it is applied, that is considered to be possessed by most⁶¹³ agents within some group. For AI, we are typically interested in achieving human common sense, which is almost always what is meant by the term. **Common Sense requires Semantic Processes to access knowledge, both to interpret messages and as a base for how new concepts are learned (that is, like all knowledge, Common Sense can constrain grounding modifications).** If a receiver does not have requisite common knowledge, including a means to form semantically similar micro and macrostates as other typical agents given messages, then they can be said to lack the associated common sense. Common Sense is considered one of the most difficult challenges for achieving AGI.

threshold). Can it be built by starting with Semantic Primitives (Wierzbicka 1996)? How are those initially formed? A hierarchical conceptual space could also be driven semantically by optimizing vector similarity (analogy-making) to find the best level to match.

⁶¹⁰ **Research: Future research should assess whether people deemed more intelligence (e.g., higher IQ or EQ scores) are able to understand deeper nestings, how such an ability correlates with Problem Solving and grounding new concepts in general, and how such nesting might be modeled in a Conceptual or Semantic Vector Space.**

⁶¹¹ "Learning" here should be generalized to evolutionary adaptation, not just lifetime learning.

⁶¹² or any environments for which the AI must re-map the grounding developed in a different training environment; this is a form of both adaption and analogy-making.

⁶¹³ "most" is not quantified. There are likely degrees to how much different knowledge, problem solving capabilities, etc. are held to be "common", where fewer common sense examples would be possessed by proportionally fewer agents. At some level, what counts as common sense would likely be debated within the group.

Integrated Information Theory (IIT)

One of the most fascinating theories related to Information in recent years is Integrated Information Theory (IIT) (Tononi, Boly, et al. 2016). IIT is a quantitative theory of Consciousness. While I make no direct claims about Consciousness in this chapter, IIT includes a formalization the authors call 'intrinsic information'. I therefore consider it in light of what I claimed is required for a system to be said to process information intrinsically.⁶¹⁴ Readers should review the extensive IIT publications, since I provide only minimalistic descriptions. While it would be constructive to compare them in more depth, I leave this for future study.

At a very high level, IIT postulates that Integrated Information, Φ , is measured as intrinsic for a system. For Φ to be greater than zero, the components of a system must have cause-effect power on the system itself, and the system must be irreducible to its components (i.e., integrated). Also, the set of components with the largest Φ value is the one that has intrinsic experience (Tononi 2015). Thus, intrinsic existence and information are associated with particular cause-effect structures.

The idea that processes must involve cause-effect properties to process information, such as the causal properties used as criteria for Φ , makes sense. However, the IIT criteria do not necessarily meet the criteria I have put forward for something to be able to process information intrinsically. I will argue that, unless IIT is an extrinsic measure for the subsymbolic processes, it is insufficient for measuring intrinsic information. I then suggest that there is potential to bring the various claims together if certain differences can be resolved.

Since this section is on how Intrinsic Information is measured, I focus on a recent paper from the IIT developers (Barbosa, et al. 2020), which introduces an information measure from the intrinsic perspective of a receiver. In essence, what is demonstrated is that information differs for a receiver fully observing information processed for a system externally (i.e., signal encoder to symbol decoder) vs. for a receiver receiving the encoded signal from within the system, such that they have no information on what was encoded and no insight into whether there was transmission error. From this, the claim is made that intrinsic information for a symbol is increased when it increases the repertoire of symbols for the receiver. On the surface, an increase in the symbol repertoire sounds consistent with my claim that a Semantic Process must form the micro and macrostates for information to then be processed intrinsically. However, there are no criteria in (Barbosa, et al. 2020) for how a message element comes to be a symbol intrinsically to the receiver. Modifying the size of the symbol set is insufficient alone for Semantic Processes. An external source piping in a new symbol with predetermined relations to inputs and outputs entails no grounding on the part of the system for which the symbol set is extended. The assumption could be that a distinction in the physical stimuli is basic enough to perform this role, but I believe this is a mistake. It assumes there is an observer already there for the distinction to be modeled as a distinction.⁶¹⁵ In fact, with predefined symbols applied by encoder and decoder

⁶¹⁴ IIT not only emphasizes the need for a theory of intrinsic information processing but also views Shannon Information as inadequate for such a theory.

⁶¹⁵ I tentatively claim that this is not an issue for Semantic Processes once reference frames are considered. The mechanisms that emerge Semantics do not need to intrinsically model the distinctions. They just need to be affected

operations, the observer can be replaced with a Chinese Room operator applying the rules of the encoder and decoder Chinese Room book. It does not matter that the observer has no extrinsic access to the encoder. The setup does not differentiate any need for grounding the symbols through the system. The observer could be replaced by a computer with a sensor that has been predefined to observe the same repertoire of symbols (or with code to instantiate symbols when certain, say, voltage differences are sensed). This leads to the major issue I have with how IIT describes intrinsic information, which is that it starts with, and therefore assumes, information by starting with symbols. In my own terms, it conflates Concepts and Percepts.

Another difference can be seen in how Barbosa et al's intrinsic information measure depends on a receiver maximizing the correct symbols received from a source. It is in this sense counterfactual, just like Shannon-like Information Theories, therefore raising issues already discussed in this chapter. For something to be considered "correct" assumes an a priori objective, and thus presupposes information itself. Intrinsic Information should be such that, rather than assuming information, it is itself required for there to be such a thing as information. Information is a difference that makes a difference to *someone*. Intrinsic information should thus be the foundation from which extrinsic information is derived. Moreover, while semantics and intrinsic information are not the same measure, the fact that the magnitude of meaning for the receiver is independent of how "correct" a reference point is⁶¹⁶ suggests that intrinsic experience is not like counterfactual information measures. My claim, of course, is that this meaning (i.e., semantics/grounding) is a pre-information phenomenon that must occur for information to be possible intrinsically. The grounding of symbols is the same as forming the symbols, and this is what intrinsic experience would require to be able to process information.

Claiming that a pre-information phenomenon is necessary to intrinsically process Information is not meant to push experience beyond the realm of Science. In fact, I propose a means to measure this, while acknowledging the challenge that the measure is necessarily performed extrinsically. More important than my particular measure is raising this general issue of how Conceptualization is inherent in describing phenomena and converts them into a different form. It is with this in mind that the measure for Semantics and criteria for Intrinsic Information I put forward do not assume the symbols we as observers conceptualize are used intrinsically by the system of interest.⁶¹⁷

(Hoel 2018) is also valuable to consider, which focuses on Emergence. At a very high level, this paper describes how causal information can increase in coarse-grained representations of systems that are multiply-realizable at a lower level. Coarse-grained macrostates of systems can

by potential differences, much like a mercury thermometer changes with temperature without processing the unit markings on its covering. The different effects can then ground stimuli such that they emerge models at a higher level that can then process them as distinctions. This is not only necessary but provides flexibility in what potential distinctions are modeled and how those distinctions are grounded as symbols, all without requiring someone external to predefine the relationships.

⁶¹⁶ This is true regardless of how "correct" might be defined.

⁶¹⁷ *Research: Consider applying the Intrinsic Difference measure in (Barbosa, et al. 2020) as a loss function during ANN training rather than Kullback-Leibler Divergence, Cross-Entropy, etc. for models where the ANN is embedded in agents. What are the advantages and disadvantages?*

be more stable because many lower-level microstates can give rise to the same macrostate. A very simple example would be temperature as a macrostate remaining steady, while the underlying gas molecules have very different configurations (i.e., microstates) and configuration sequences. To the extent that the multiple realizations at the lower level make no causal difference to some agent, it is beneficial for that agent to perceive the system at a coarse-grained level. I have claimed that perceiving phenomena at some emergent level is heuristic. (Hoel 2018) goes further, providing insight into why it is valuable to an observing agent to experience stimuli at a coarse-grained emergent level.

A question is whether this in any way helps to understand how information can be intrinsic. My claim is that the causal emergence measure provided is itself for *extrinsic* emergence.⁶¹⁸ This certainly may be as intended by the author, as “intrinsic” is never even mentioned. My concerns are therefore not directed at any claims made in this paper. However, IIT generally does point to conceptualized causal properties like these as addressing intrinsic information and experience.⁶¹⁹ My claim is that such properties provide no obvious insight into how a given agent intrinsically emerges these extrinsic emergent phenomena. My rationale is similar to that provided above for (Barbosa, et al. 2020), so I will not repeat them here. Instead, I add a related speculation from the perspective of the earlier Emergence chapter. There, I stated that one is tempted to assume that the Mind is also Emergent, whereas I claim that the Mind is better understood as Emergence. If one *starts* with the assumption that the Mind is Emergent, then an explanation for how any external system becomes perceived as emergent can seem like a separate example from how the mind is also emergent (e.g., how H₂O emerges as water). However, this conceives of the mind in the third-person, just as for those other systems. One then must describe how *it* is perceived in an emergent manner, just like those other systems. Instead, if one understands that it is the Mind of the receiver that constructs the emergent properties for any of these external systems, including a conceptual form of itself, then we can at least avoid conflating an explanation for extrinsic with intrinsic emergence. IIT seems to provide an insufficient notion of intrinsic information because it assumes that the Conceptual (symbolic, intentional) description is the same as the intrinsic phenomena it conceptualizes. If so, then this leads to an unrealized⁶²⁰ conversion for what is described into an extrinsic emergent form that has no clear relation to the intrinsic emergence that the system itself may have. When that system of interest is our own experiences, then the concepts formed fail to explain that which is being conceptualized, and therefore that which is of interest (i.e., our own intrinsic experience). Percepts are neither

⁶¹⁸ Of course, the idea of an agent is an abstraction that is multiply-realizable. However, this has no obvious relevance to the claims in this section. The concept of an agent is an extrinsic emergent projection, no different in kind than other extrinsic emergent concepts.

⁶¹⁹ (Hoel 2018) does not even reference IIT, so I stress again that I only use the paper to illustrate issues with IIT claims, not this specific paper.

⁶²⁰ Although perhaps only a terminological difference, (Tononi 2015) describes an experience as a conceptual structure. I claim that intrinsic experience is at its core non-conceptual. I am making an assumption here that the authors assume that experience must be intentional and have not considered the role of reference frames in distinguishing symbolic from subsymbolic processes.

fundamentally symbolic nor intentional.⁶²¹ A challenge with a Causal approach to intrinsic information is if the system of interest is assumed to experience the same symbols as the set involved in the causal relations that process them. Such a model would fall into the ERM category described earlier, and thus suffer from the same issues.⁶²²

Despite the above concerns, the formalizations for Causal Emergence provide a very sensible measure of the functional value of emergence from an external perspective. They address a specific and key role for causality, which I have in no way addressed. The same can be said for IIT more generally. However, there is another potential way where the value of IIT's formalizations can be used to provide an explanation for intrinsic information processing. The intrinsic information criteria issues that I raised for IIT go away to some degree if the dynamics by which IIT is associated with intrinsic information occur at the subsymbolic level. That is, if measured at a level where the symbols modeled are considered subsymbolic (i.e., not the symbols for some system of interest's own reference frame), then IIT might provide a measure of semantic-ness for some theoretical system of interest it could emerge at a different level (i.e., a reference frame from which the processes IIT is measured against are subsymbolic). Given that IIT examples are often measured over neuronal networks, it is a fair assumption that this is what they have in mind. (Tononi 2015) proposes that a geometric Cause-Effect Space determines the quality of experiences. This emerges from the self-causal dynamics of the underlying network, with no necessary correlation with any labels added by external modelers to the individual nodes and edges. Thus, as subsymbolic dynamics, IIT could help explain what mechanisms that emerge percepts look like extrinsically when converted from subsymbols to symbols (i.e., changing reference frames). Perhaps IIT could then serve as a semantic/grounding measure.⁶²³ If IIT claims are necessary or sufficient for measuring how a subsymbolic level grounds an emergent level from a different reference frame, then there is no major conflict with my criteria. My claims would simply provide additional understanding of, or perspective for, grounding/semantics and intrinsic information processing.

There are some potential tensions that would still need to be worked out to integrate my claims with this understanding of what IIT measures. One difference is that Semantic Processes, while potentially causal, are not a set of counterfactual causal inference-based calculations. This is important for it to be a measure of something fully intrinsic. A noncounterfactual intrinsic measure is not only natural but more tractable (e.g., no need to sample/perturb to achieve counterfactuals for determining a causal probability distribution) vs. agents fundamentally performing formal inference to generate their own semantic and intrinsic information

⁶²¹ As I have stated, Percepts can form conceptual properties but do not have to. These are particular perceptual experiences that include a conceptualization of some phenomena. Even in these cases, the intrinsic emergent experience is foundational as the medium within which properties of conceptualized phenomena are experienced.

⁶²² Also consider that, while we tend to assign intrinsic emergence at the natural language level for ourselves, there is no complete causal model at that level that determines what words are generated. For the Semantic Processes I have described, the underlying processes (e.g., neural dynamics) select macrostates that generate the emergent symbols.

⁶²³ IIT measuring grounding makes sense if it is a measure for consciousness, since the two should be related. However, even if IIT did not measure consciousness, it could still potentially measure grounding.

properties.^{624,625} While there is no reason to believe IIT is intended to claim that the systems with $\max(\Phi)$ measure their own Φ , let alone comparing it to other possible Φ 's, it does mean that Φ , if measured counterfactually, must be an extrinsic model *about*, rather than *of*, an intrinsic process. It thus conceptualizes what is a non-conceptual phenomenon. According to my claims, this converts the processes into eliminable symbols for which information can be measured as extrinsically projected onto the system itself. Semantic Processes, as I describe them, are noncounterfactual phenomena that give rise to such a symbolic structure. It is meant to measure the functional equivalent of the richness of the experience from the perspective of the system of interest. I have claimed that any symbolic measure, whether Shannon Information or IIT, does not directly represent the experience of a receiver without this Semantic Grounding, since it is a conceptualized extrinsic macrostate that is not symbolic to the system of interest without Semantic Grounding to make it symbolic *to that system*.

Another difference is the claim that a Von Neumann computer architecture, as well as any feedforward network, has no capacity for Integrated Information (Tononi 2015). My claims do not dismiss that either of these could ground symbols. I suggest that there would be no way to prove that they do, but that this is true for any actual system (or at the very least for any actual digital system, in the off chance that new physical mechanisms are discovered that completely change our understanding of mechanisms). Based on my reasons, it is unclear why Semantic Vector Space models built on Von Neumann machines, many of which are built on purely feedforward NNs, are at any *fundamental* disadvantage compared to other actual systems.⁶²⁶ This requirement by IIT raises some issues according to (Doerig, et al. 2019), which claims that IIT⁶²⁷ must be false or beyond the realm of Science, because the feedforward network is a universal function approximator and thus could approximate any structure that IIT would claim has nonzero Φ .⁶²⁸ One could simply claim that this is just the way it is: Φ is independent of the

⁶²⁴ Claims that the universe is fundamentally information (Lloyd 2006, Davies and Gregersen 2014) have become increasingly popular. I have claimed that Information makes no sense without an observer, and that Information requires that Semantics first form the micro and macrostates against which Information is actualized. Claiming that Reality is fundamentally information therefore not only first requires an observer to somehow be equally fundamental but must also explain away the essential role for Semantic Processes. Information alone cannot explain the emergent phenomena we experience or how something can be informative. It is difficult to see how it could be experienced as about something else if all something else's are also information. One can of course model the Universe's computational capacity, since one can do this for anything modeled (e.g., ERMs). However, this first requires conceptualizing the Universe, which assumes an observer to form the conceptualized model. In that form, the Information associated with the model is projected extrinsically onto the system.

⁶²⁵ Note that the issue with intractability is not just a practical problem if the goal is to identify something measured intrinsically (e.g., for a human). I assume no one is suggesting that system's explicitly measure their own IIT, but this raises questions about what information it is processing intrinsically that would be measured. The extrinsic measure should differ from the intrinsic measure.

⁶²⁶ *Research: Perhaps the IIT criteria for consciousness are more restrictive than for grounding (of course, I claim that grounding is a prerequisite criterion for conscious experience, so grounding cannot be more restrictive than consciousness criteria)?*

⁶²⁷ The claims are made for any causal structure-based theory of consciousness.

⁶²⁸ Any recurrent network with nonzero Φ can be represented as a feedforward network by unfolding it in time (LeCun, Bengio and Hinton 2015).

type of functions that systems implement. However, this then raises questions about what exactly IIT explains. If the processes that systems perform on inputs do not matter, then the only justification for IIT I can think of is because biological agents, such as humans, have recurrent networks and are conscious. However, without any justification other than this association, it is equally justified to suggest that any neurological property explains consciousness (e.g., particular neurotransmitter balances, electric field properties, microtubule properties). The criteria for Semantic Processes raise no unique issues for feedforward networks to in theory process information intrinsically.⁶²⁹ Even so, I am open to more restrictive criteria for grounding. I leave it as an open issue as to whether feedforward digital NNs can really ground symbols or even approximate true semantic processes in the limit as I have proposed. More broadly, my hope is that these differences can be addressed to provide a more comprehensive understanding of grounding and intrinsic information processing.^{630,631}

Summarizing the Relation between the Processes

In this chapter, I have spent a considerable amount of time working through Weaver's three information types. I referenced Semantics and Effectiveness as Processes, rather than Information proper to focus on their core, and most difficult to define, properties. I have restricted the term 'Symbol Grounding' or just 'Grounding' to being a semantic function.

Given my descriptions and claims for the three distinct processes, the following relationships are made: **Semantic Processes determine indices for Shannon-like Information Processes over models, which describe how the models formed are updated given the indices. Effective Processes constrains what indices are formed.** Semantic Processes do not alone explain *why* certain properties are associated with certain symbols. Effective Processes determine what heuristics are adapted or learned, while Semantic Processes describe the heuristics themselves⁶³². Effective Processes are associated with what grounding emerges but does not measure the grounding itself. There are infinite ways to describe a situation. This is what gives

⁶²⁹ There are practical issues with feedforward NNs. Feedback loops between neurons are found extensively in biological brains and it would be very surprising to discover that they are not necessary for AGI. These and other biological properties not modeled in most ANNs have been discussed in a variety of publications and will also be discussed in a future manuscript.

⁶³⁰ *Research: Consider how IIT could be extended to measure causality and irreducibility with respect to a Conceptual / Semantic Vector Space based on how they define the properties of a Cause-Effect space.*

⁶³¹ *Research: Explore ways to unite this manuscript's claims with IIT, including the distinction between extrinsic and intrinsic emergence, Percepts vs. Concepts, and the relation between Cause-Effect Space and Conceptual / Semantic Vector Space. Is there a potential relation between Semantic Specificity within a Conceptual Space and $\max(\Phi)$ in a Cause-Effect Space? How correlated are their geometric forms? Could Cause-Effect space be used to determine the quality axes for the Conceptual Space (e.g., the forms in Conceptual Space are "holographic" to the Cause Effect Space dimensions)? A starting point for contrasting it to Semantics as Grounding would be measuring Φ for ERMs. If Semantics describes processes that make information available intrinsically, then IIT should account for the fact that the semantics in such models are parasitic on the human.*

⁶³² More accurately, Semantic Processes are associated with experienced heuristics. Affordances, on the other hand, are associated with both experienced and unexperienced (or at least not directly experienced) heuristics (e.g., various motor reflexes). Also keep in mind that both Semantic Processes and Affordances depend on the receiver but are measured for the message / environment state received.

rise to the Frame Problem (Shanahan 2016). Somehow, we can use semantics to coordinate effectively. Semantic Processes are largely determined by their Effectiveness (i.e., how the grounding for a message helps cope with the world). Because Semantic Processes could be constrained in countless ways, being able to measure grounding for AI agents is important for them to emerge semantics similar to our own given similar Effective constraints.⁶³³ At this stage, I cannot imagine a viable option for achieving AGI capabilities, such as Common Sense, without accounting for these processes. A better understanding of some of our own effective constraints may first be necessary, while other insights will be gained through the research itself. Thus, the research should be interdisciplinary.

From an Information Theoretic perspective, a heuristic decreases effective uncertainty at the expense of not always obtaining. **Due to our physical limitations, we represent phenomena in the world as emergent, heuristic concepts and concept instances. Heuristic interpretations of the world are experienced as Semantics. Effective Processes shape what semantics are adapted. That is, semantic representations are shaped (i.e., constrained) by their effectiveness with respect to implicit or explicit goals.**^{634,635,636} Semantics itself is likely a heuristic adapted for its evolutionary effectiveness, and therefore, in a sense, a meta-heuristic. Semantics not only enable Shannon-like Information processing but also enable us to communicate with reduced resources (i.e., message lengths) both internally (i.e., thinking) and to others by leveraging shared grounding in our respective Knowledge Systems. Semantics was adapted for the heuristic value of emergent representations that would otherwise require forming models requiring more information capacity without sufficient additional benefit.

I use Chess as an example to summarize the relationships between the Shannon-like, Effective, and Semantic Processes. Like the real world (but simpler), chess includes constraints, such as how different chess pieces can move and a goal that drives how the chess pieces tend to behave. Assuming the standard objective for chess, one could say that a player **effectively** constrains how symbols are **semantically** grounded if those symbols (e.g., a chess game state) are processed in a way that enables chess playing such that the probability of winning is increased (or leads to winning). An effective grounding could be learned by playing many previous games, potentially along with learned or innate propensities that have been developed in other ways that can be applied to Chess. Other goals are possible, such as learning, so as to improve future chances of winning. This could change the grounding constraints and therefore the grounding

⁶³³ **Research: Research the properties of Effective Processes as Semantic constraints.**

⁶³⁴ Note that Effectiveness is not the same as one's stated goals. The explicit experience of messages impacting one's stated goals is a semantic property (i.e., interpretation of messages given one's semantic *concept* of goals). Effectiveness can still be measured with respect to any chosen goal, since the act of choosing a goal is the act of an external measurer of some system *for the purpose of measuring* Effectiveness. This differs from the semantics the receiver happens to have that maps to goal concepts stored in their Knowledge System and how that is impacted given a received message.

⁶³⁵ I am not claiming that goals, or how they are currently modeled in AI, are sufficient for measuring effectiveness for AGI. However, effectiveness can in theory be measured against any goals of interest to the AI user.

⁶³⁶ **Research: Consider how context and attention mechanisms can be used to focus on certain, say, domains in a semantic vector space, and how they can optimally be shaped by effectiveness themselves.**

(although the constraints may be similar for Chess, given how bounded the game is). Effective constraints on grounding requires that the different symbols (e.g., discrete classes for types of chess pieces) are sufficiently separated so they are effectively associated with different (allowable) chess moves and chess strategies given different observed game states.⁶³⁷ One's implicit or explicit objective(s) can constrain grounding properties, which should be measurable as an Effective Process.

Semantics should likewise select **Shannon-like** models such that, given certain environment properties, the model enables effective behaviors on the part of the observer/player. **The micro and macrostates cannot be arbitrarily formed since the potential for information must be available given the environment to be effective. Indeed, this may underlie much of the extrinsic emergence properties of complex systems: Grounding across interdependent concepts should be semantically consistent to create stable, related abstractions across them, such that grounding can be used for generalizable inference, such as analogy making. Again, this leads naturally to Conceptual Spaces.** Going back to the Chess example, one might use their grounding of wars to assess attrition risks, grounding of individual chess pieces to determine possible moves, or grounding a pattern over a few different pieces to consider a coordinated tactic.

Given a message, such as symbols representing a game state (e.g., configuration of current chess pieces), Semantics actuates Shannon-like Information as an interpretation that updates some model the player holds about the game. This actuation is measured with respect to the probability of features of the message, which is the **Shannon-like Information** of the message. The features can represent discrete concepts at any level, including the opponent's next move conditioned on what move we make. At certain levels, such as interpreting game states with respect to the opponent's strategy, Semantic Information Theories can be applied. All of these information theories relate to uncertainty with respect to an underlying microstate, which could be the set of winning states, or other states, such as uncertainty about the microstate that determines what moves will be made in different states. Each new game state can generate Shannon-like Information.⁶³⁸

These claimed properties for Semantic and Effective Processes point back to a potential limitation discussed in the Performative Systems chapter for how embedding methods currently build Semantic Vector Spaces. **The way embeddings are trained may capture or at least approximate some Semantic Processes but without any direct dependence on Effectiveness.** This may work reasonably well in certain cases.⁶³⁹ However, the optimization functions used over

⁶³⁷ *Research: Determine if/how the size or openness of the actions and transitions in a domain limit potential grounding for agents isolated within that domain.*

⁶³⁸ *Research: Shannon Equivocation measures the uncertainty that the message received is the one that was meant to be transmitted. Consider how Equivocation can be applied more generally to intrinsic model noise, such as loss from data represented as emergent features and missing causal or latent variables.*

⁶³⁹ As a theoretical example, consider a set of grammatical rules that constrains how certain words are used may have been selected due to its effectiveness. This could then produce regularities in the words observed that can be used to learn to ground them. In this case, the use of word conditional frequencies in data may act as a decent surrogate for their effectiveness but in limited domains. Even in those domains, they will be imperfect, especially for "edge cases".

data (e.g., cross-entropy given word predictions) do not directly optimize for effectiveness similar to what shaped their grounding in humans. Learning in ways closer to biologically intelligent systems, such as humans, should require an objective function(s) that optimizes Semantics based on Effectiveness, which in turn requires data that can be used to sufficiently assess Effectiveness of behaviors driven by Semantic Processes.⁶⁴⁰ This likely requires not only multiple sensory modalities, but the ability to act in the world within which those modalities become available.

The relations between these processes are important. Semantic Processes have a mutual dependence on Shannon-like Information because Shannon Information is enabled by Semantics Processes, and Shannon Information then evokes new forms within the Semantic Knowledge System of a recipient. This in turn influences what new micro and macrostate are formed and updated. Similarly, Semantic and Effective Processes have a mutual dependence. Effectiveness shapes what Semantics are possible and their propensities given receipt of different messages and contexts⁶⁴¹, while the Semantics evoked then shape effectiveness that constrains future semantic model formations.⁶⁴²

Conclusions

I began this chapter by claiming that a system must have an observer to process Information intrinsically. I then proposed that Semantic Processes form the micro and macrostates against which information is measured for messages, and thus fulfill the role for a system to intrinsically process information. Semantic Processes are therefore best understood as pre-Information phenomena. They interpret messages in the sense of determining how existing knowledge is informed given the message received. Given a message, Semantic Processes could be conceptualized as constructing the symbol indices that form Shannon-like micro and macrostates.⁶⁴³

Since all Information Theories are based on updating models with symbols, there is little benefit in expanding them to cover subsymbolic processes. This preserves what Information means mathematically and leaves existing Semantic Information theories intact. Defining the

⁶⁴⁰ *Research: Consider how this claim relates, extends, or is extended by Hoffman's Fitness-Beats-Truth theory (Hoffman 2019).*

⁶⁴¹ Technically, context should be considered part of messages but is used here to emphasize that a receiver's internal state affects the symbols grounded for messages.

⁶⁴² *Research: To measure and relate Effective, Semantic, and Shannon Processes, apply them to evolve agents to emerge their own Semantic processes. This does not guarantee that the agents would evolve/learn the same implicit mapping between messages and models but is a good research area to determine how the models might be constrained to do so. This requires constraining the agents to coordinate via communication but with less bandwidth available than needed to fully describe events or objects at the explicit feature level in the environment that must be communicated to achieve some objective, thus requiring lossy abstractions to emerge for messages through use of agents' shared knowledge.*

⁶⁴³ *Research: How does a micro and macrostate index relate to the concept of an Index in Semiotics? The latter do seem to be about something, and thus seem intentional. Note, however, that Semantic Processes determine indices; they are not equivalent to the indices themselves. Semantic Processes are required to develop conceptual (i.e., intentional properties) for given perceptual grounding. The use of the term index in Shannon-like equations is specific to what symbols are contained in the set applied. This seems different than Index as a type of Sign. However, it is worth considering how the various claims in this manuscript relate to non-Symbol Signs (Indices and Icons).*

fundamental role for Semantics at this subsymbolic level means no Semantic Information Theories must be the “right” one. They all assume different reference frames, the utility of which depends on what one wants to measure. A corollary is that any Information Theory depends on Semantic Processes. I even claimed that it makes no sense to call something a “symbol” for the system modeled as receiving it without the underlying Semantic Processes to determine what they symbolize. Specifically, we must not conflate what is a symbol to us as semantic observers vs. what is a symbol to the system that evokes the symbol in us.

I proposed that Conceptual Spaces offer a model type that meets the criteria for Semantic Processes. This should enable us to carry forward all the existing properties and formalizations for them as a means to model systems with intrinsic/semantic processes. Semantic Vector Spaces can approximate Conceptual Spaces, offering a way to build actual models that at least simulate Semantic Processes. However, there is a need to constrain them with Effective Processes.⁶⁴⁴

The most difficult challenge faced in this chapter was to propose a way to describe grounding properties while staying true to the need for Semantics to be explained intrinsically. Here my claims were more tentative. While I am confident in defining subsymbolic phenomena as outside of and, at least conceptually speaking, forming the micro and macrostate symbol set, there is no clear way to determine how the subsymbol / symbol distinction could be proved in actual systems. The issue is that the processing of any realized system can always be causally determined by fundamental mechanisms (e.g., atoms, bits).

I suggested that the difference between Semantic Processes and Shannon-like Information directly relates to the distinction I made in earlier chapters between Percepts and Concepts. **It is only conceptualizations of macrostates, including the microstate/macrostate formation process, that is covered by Shannon-like Information. As conceptualization is often necessary to describe any phenomena, their description necessarily gets enveloped in terms amenable to Shannon-like formulation. We might say that Percepts are Semantic, and Conceptualization is the process of “Shannon-izing” those semantic properties. Just as Concepts have perceptual properties about percepts, Shannon-like symbols have semantic properties about the semantic properties that ground them. Semantics are fundamental.** This clarifies why it may seem like Semantic Processes might be explicitly and counterfactually measured like Shannon Information, as is done with current Semantic Information Theories.⁶⁴⁵ Communication necessarily relies on

⁶⁴⁴ These subsymbolic criteria have implications for explainability in AI. Just as for biological agents, grounding in AI must form the explicit symbols as emergent heuristics, which are often lossy. There is a Description Gap due to the irreversibility between the underlying dynamics and the symbols that emerge. Humans have adapted communication and Theory of Mind processes to gain mutual understanding, which relies on common evolutionary and experiential constraints. If AGI were to achieve the same constraints, then we would likely need to use similar processes to trust them. Heuristic techniques we have adapted to find Common Ground may be useful in this regard. AI may provide additional means to gain understanding of outputs since we have better access to their underlying dynamics than we do for biological agents. However, the criteria I have claimed for grounding/semantics come with a necessary degree of irreversibility. While human understanding and trust is incredibly important for AI, it is also important to not throw out the proverbial baby with the bathwater by forcing AI to be modeled more like ERM.

⁶⁴⁵ Other Information models like Algorithmic Information Theory (Chaitin 1969, Hutter 2005, Kolmogorov 1963) and the related Solomonoff Induction (Solomonoff 1964a, Solomonoff 1964b), are also grouped into Shannon-like Information here. They require updating a set of predetermined model variables and relationships given new data.

these symbols, making them seem all encompassing. It is only the *concepts* of Semantic Processes that can be symbolized, which converts them into something else that does not explain how any symbols become semantically grounded. A Concept differs from the Percept within which it is contained. Again, language itself puts us at a disadvantage for understanding the distinction I make, because it necessarily encodes semantics into symbols, and thus concepts. As modelers, we can decode such symbols to obtain their semantics, but there is inherent difficulty in trying to describe this process itself. This brings us back to what underlies the challenge in defining criteria for systems to ground symbols and intrinsically process information. Any extrinsic description converts phenomena into Concepts, and Concepts are subject to eliminative materialism.

While difficult, this may not be impossible to define and even measure. I proposed a way to envision what a theoretical model that could process intrinsic information must involve and suggested that one can then empirically assess whether certain realized systems seem to fit the criteria provided. Grounding is not epiphenomenal, as can be seen in the differences between ERMs and ANNs described at length in earlier chapters. The tentative criteria proposed do not fully escape the use of extrinsic assessment but at least constrain its use in such a way to distinguish models that are and are not parasitic on the use of an external system's grounding. That is, the proposed measure is extrinsic but does not resort to assuming the same properties extrinsically projected onto the systems of interest are also possessed for those systems. Thus, the definitions I provided in this chapter attempt to capture what intrinsic information might look like from the external point-of-view, which differs from a blanket assumption that extrinsic emergent phenomena provide appropriate representations of intrinsic information processing.

Conceptual Spaces are any ideal representation for this theoretical model. They have continuous subsymbolic capacities to form an infinite number of percepts and concepts with semantically meaningful distances between each other based on their vector values along quality dimensions. ANNs can produce Conceptual Space-like models (i.e., Semantic Vector Spaces) through embedding methods. Semantic Vector Spaces might merely approximate the Semantic Processes criteria that Conceptual Spaces achieve. However, the criteria proposed are such that I can speculate that they are equivalent to Conceptual Spaces "in the limit". This requires more specific and testable hypotheses.

I hope these claims lead to better theories and modeling for AI and perhaps even in an understanding of ourselves and our own knowledge.⁶⁴⁶ Semantics is the Reality we as sentient

This holds even for distinctions as basic as 0s and 1s. As I said earlier, there must be an observer for which distinctions are modeled as distinctions for there to be a model intrinsically to update. As for Shannon Information, however, they can potentially be applied at what is assumed to be a subsymbolic level and emerge semantics at another reference frame.

⁶⁴⁶ To point to just two areas where understanding the processes in this chapter may help, AI struggles in applying Context and in Adaptation compared to humans. Both can be improved using Semantic and Effective processes. Context-driven understanding is largely implicit. There is no explanatory layer between neural processing and natural language explanations, due to the Description Gap. (Zimmermann 2015) provides some intuition about the centrality of context and a form of adaptation in mental processes: "...the meaning of a particular statement depends on the larger context: a whole within which the part has meaning. It may be less obvious that we usually intuit such a larger

intelligent agents experience. It is difficult to imagine how AGI could do without it. Future work is required to determine if the claims made are correct and what else is needed to better understand and advance the phenomena they are meant to describe. Also, this chapter barely addresses the wealth of thought and research in Semantics. At this stage, I can only hope to have added new considerations, or at least point AI research in their direction through increased awareness on the limits of symbol-level information as a starting point for modeling intelligence.

To end on a philosophical note, even though the criteria I provide for grounding cannot be used to prove intrinsic emergence, it seems plausible that they are relevant for how intrinsic emergence occurs. I mentioned that a Conceptual Space, from the reference frame of the model itself, seems better defined as a Perceptual Space that includes Concepts. Grounding involves dynamics that form but are outside the symbols grounded within this Perceptual Space.⁶⁴⁷ A challenge for theories on Consciousness is that they start with Concepts. While this is true for all of Science and Philosophy, it does not create complete failure in most Sciences because Concepts can be used to evoke the common semantic ground, and thus produce perceptual knowledge.⁶⁴⁸ However, the Percept/Concept distinction strikes at the very heart of what Consciousness is. Here, we must explain the very issue limiting explanation. The Semantic / Shannon-like distinction seems intimately related to the fact that we all know exactly what consciousness is like but cannot come up with even a candidate scientific theory that would address the Hard Problem (Chalmers 1996). The (symbolic) *concept* of Qualia is eliminable, while the (semantic) *percept* for the experience is not.⁶⁴⁹ A fundamental role for Reference Frames in this divide may help unify some perspectives on the matter.

context based on trained expectations. Let's say, for example, you pick up Shakespeare's Hamlet for the first time. Looking at the book, you probably know that it's a play of the kind we call tragedy and contains the line 'to be or not to be: that is the question'. You have heard this line many times before, in sitcoms or even advertisements. All these prior influences shape your expectations of this play in some way. The hermeneutic circle means that some greater context always influences how we understand a particular part. As you read, however, you will have to revise not only your view of what the play is about but also what this famous line actually means. The point is that whole and part influence each other. Our better understanding of particular lines will reshape our grasp of the whole."

⁶⁴⁷ More specifically, one can have symbols for those stimuli, but these are explicit conceptualizations of them, which are not the same as the stimuli themselves. As an example, my conception of neurons and their dynamics differs from the neuronal dynamics that emerge those concepts.

⁶⁴⁸ It was not an aim of this manuscript to determine if the Sciences are impaired by the need to use conceptualized phenomena. Perhaps this is true. However, for the Sciences and other areas of knowledge, we can at least evoke and leverage our own grounding capabilities as we share concepts in symbolic form with each other and ourselves. It is a different story for building AGI. Because AGI is aimed at building the very thing that produces concepts, it cannot just be a model of concepts. It must have the underlying perceptual grounding space to form concepts like all biological intelligence we know of does. This is required to be a modeler. This also points to what ANNs are showing us; something fundamental that is missed by ERMs, which is the ability to form percepts, or at least some functional approximation of this ability. With this understanding, there is potential value in AI researchers considering where "*perceiving*" is done and in what form for ANNs. Also important is to determine if they can form Concepts from their perceptual base, and if not, how we enable them to, based on any possible insights for how biological agents do it.

⁶⁴⁹ I have to wonder if the fact the scientific theories end up with consciousness as epiphenomenal should be showing us a more fundamental issue with how we go about explaining/inquiring about phenomena. The Percept/Concept distinction perhaps is pointing to this larger issue. Any conceptual account of Consciousness, let alone a functional

account, leaves out what it is meant to explain. I am unsure how to construct a nonconceptual theory or if it is even possible. That this is so difficult demonstrates both the power and limits of Concepts.

Chapter 5: Conclusions and Future Research

Conclusions

Some claims in the manuscript are provided with more conviction than others. For example, I am confident in the claims about Extrinsic and Intrinsic Emergence, including how they are and are not subject to Eliminative Materialism, respectively. I am also confident that the Percept/Concept distinction sheds light on this issue and provides a way to understand what causes the Hard Problem of Consciousness. I am confident that understanding the role of Reference Frames as fundamental to these distinctions helps understand them as well. I am confident that there is no intrinsic emergence with Symbols or Information only. Experience, at any level, requires Semantics to form the grounded for symbols. Without any grounding, there are no symbols, no observer, and no intrinsic information processing. Finally, I am confident in the criteria put forward using Reference Frames to assess whether these properties would exist for some theoretical model. Where I am less confident is in the empirical assessment proposed for actual systems, whether digital or physical. I am even less confident that empirical tests can be formulated that would clearly fulfill this role. Even if they could, they might not be straightforward to measure in actual models. Candidate measures for Semantic Processes should also only be considered notional. While the more confident claims do not rely on the less confident ones, they are generally associated with how to actually measure Semantics/Grounding, which is assumed to be important to measure to achieve AGI. My hope is that the more fundamental claims help clarify what must be measured, thereby facilitating and focusing discussion and research efforts. With this in mind, the main conclusions of this manuscript are:

Chapter 1: Emergence and Grounding have critical functional roles in biological intelligence, and seem equally critical to AI, particularly AGI. A Description Gap limits the ability for lower-level mechanisms to explain or explicitly map to emergent phenomena. Implicit grounding is the only way to connect explanatory levels. When observing or modeling systems, ‘Extrinsic Emergence’, which is emergent properties about external systems but experienced outside those systems, is easily conflated with ‘Intrinsic Emergence’, which is experienced by the agent. It is only when we conceptualize emergent experiences, which converts them into extrinsic emergent phenomena, that they are subject to Eliminative Materialism. Concepts are Percepts with properties *about* phenomena. Percepts are fundamental. Concepts differ from their underlying perceptual form, and thus cannot explain how they are evoked. The fact that we must reason and communicate at a conceptual level to understand and discuss nonconceptual content relates to the Mind-Body Problem and the challenge of *explaining* the Mind in relation to AI. Consciousness (i.e., The Mind) is not Emergent, it is Emergence. It is therefore uninformative to claim that consciousness is like other emergent properties.

Chapter 2: If Artificial Neural Network (ANNs) models can ground symbols, they would not currently ground them anything like how biological agents would ground them. This is in part due to how they are trained. Our own grounding is primarily shaped by how it enables coping with the world. Indeed, this ‘performative’ aspect of agent reasoning should drive the

‘representational’ aspect that so much of AI research has focused on. This focus on representation and explicit truths is even stronger in Symbolic AI. AI models must have rich and frequent interactions shape how they ground their environment as affordances. This is necessary to achieve what is fundamentally a performative intelligence, where observers are actively involved in events and objects as they unfold to them. How effectiveness grounds our percepts and concepts cannot be modeled explicitly or trained in without rich, complex interactions.

Chapter 3: Many AI modeling approaches, which I call ‘Explicit Representational Models’ (ERMs), are not grounded. Like text in a book, ERMs evoke extrinsic emergence in observers but do not themselves ground the concepts. This is due to the way information is modeled for them. ERMs are Models, not Modelers. AGI, like biological intelligence, must be able to model, not just be a model. This limits the potential performance of such models. It is only through implicit processes that AGI can model effective relationships to lower-level mechanisms in the world. For this, AI must learn to form its own symbols. Unless these symbols map to subsymbolic processes effectively, systems we design cannot meet the expectations for AGI, including their ability to predict or infer the properties of phenomena about which they would be expected to reason. How this implicit mapping is represented in models is of critical importance. Conceptual Spaces seem like the most appropriate way to represent such knowledge. Semantic Vector Spaces appear to at least approximate Conceptual Spaces. If so, they provide a promising path to build them through learning and adaptation via Neural Embedding techniques.

Chapter 4: The information processing we attribute to AI models is based on our own extrinsic emergent projections, which do not necessarily bare a relation to intrinsic information processing. Information Theories provide no criteria for a model to itself process the information it receives. For this, Grounding, or what I call ‘Semantic Processes’, is required. Semantics Processes are pre-information. They determine the micro and macrostates, against which information is measured. This understanding of Semantics differs from existing Semantic Information Theories. These theories assume symbols and their meaning rather than explain how the meaning is achieved. As symbolic measures, they do not address the critical role of Grounding as a subsymbolic process to form symbols. Unlike ERMs, a neural network is *in theory* subsymbolic to the Conceptual Space it forms. It does not matter what is symbolic to us from our external reference frame. Models that meet the subsymbolic criteria from their own theoretical reference frame can ground symbols and process information intrinsically. This associates the Grounding/Information distinction with the Percept/Concept distinction. Note that the way Semantic Vector Space models are currently developed may capture or approximate certain Semantic Processes but without any direct dependence on their Effectiveness. While Semantic Processes determine the micro and macrostates formed, Effective Processes shape the Semantic Processes learned/adapted and thus constrain what micro and macrostates form.

The most wide-ranging claim may be providing a fundamental role for Reference Frames and their relation to how we conceptualize percepts. It has potential implications for a variety of theoretical issues, perhaps even for paradoxes, associated with the computational nature of AI. I suspect that few will be excited about the idea of incorporating a frame-relativity role into our

understanding of the nature of computation, information, and cognition. However, I feel compelled to put it out there for serious consideration.

Future Research

As said near the beginning of this manuscript, current AI, including Deep Learning, is not on a clear path to AGI. At the very least, Neural Network architectures and how they are trained will require major changes. However, even if core aspects of how these models are built, infer, and learn must be rethought, they have demonstrated something I believe will ultimately be a part of the AGI story. We need to better understand and develop it. My hope is that the perspectives and connections made in this manuscript help focus research investments along these lines, including studies, formalizations, and new modeling approaches. For AI researchers and investors, I offer the following: You are busy and have limited funds: I get it. You may be externally motivated to solve specific problems, and it is a challenge to improve on what someone else just published. You have profit goals that you need to meet to stay relevant in your field or business. However, many of you were driven to this technical field because of either the excitement and importance of building AGI or to solve major problems that, if we are honest with ourselves, requires AI capabilities that are nowhere in sight. We need to bring the focus back to where it ultimately matters most and where most of us find the true motivation that we stifle with false hopes for short-term convenience. This will require a significant cultural shift in science, government, and business, because none of us can do it alone.

Many important topics associated with Intelligence were not addressed in this manuscript. Those that were require integration with existing theories and further development. A separate manuscript is planned to address how a variety of equally important areas connect to the claims made here and by others. Topics will include Causality, different types of Signs, different modes of Inference⁶⁵⁰, Prediction, Reasoning, Common Sense, Stigmergy, Complex Systems, and Neuroscience⁶⁵¹. It will also discuss a body of research outside AI that seems particularly promising to help develop AI on more solid foundations, including without limiting it to narrow predefined objective functions (e.g., reward, fitness, explicit goals). Finally, I hope to lay out candidate research programs for achieving AGI.

Even if correct, this manuscript is clearly not the whole story. Much more remains undiscovered to achieve AGI and understand our own cognition.⁶⁵² In parallel with modeling and

⁶⁵⁰ Deductive, Inductive, and Abductive Inference are the established modes.

⁶⁵¹ This includes how the Description Gap and Complexity relate to understanding the Brain through Neuroscience and Cognitive Science, and how it suggests we implement the many insights from biological fields as “guideposts” to advance AI capabilities.

⁶⁵² *Research: For example, Neuroscientific studies can help find more promising or specific ways in which Conceptual Spaces must be modeled. They might also help understand how biological intelligence structures and applies reference frames to leverage emergent dynamics effectively. This includes what the neurophysiological equivalent is to “forming micro and macrostates”, and accordingly, how concepts arise from percepts. Of course, it would not be surprising to find that more realistic neuronal dynamics must be modeled in ANNs to provide effective models capable of these processes.*

engineering efforts, this will require that researchers have the opportunity to commit to go as far back as needed to get there. We will learn a lot about ourselves along the way.

References

- Ackerman, E. 2014. "Can Winograd schemas replace Turing test for defining human-level AI?" *IEEE Spectrum*.
- Anderson, P W. 1972. "More is different: Broken symmetry and the nature of the hierarchical structure of science." *Science*, 177 (4047) 393–396.
- Ashby, W R. 1958. "Requisite Variety and its implications for the control of complex systems." *Cybernetica (Namur)* 1:2.
- n.d. *Autoencoder*. <https://en.wikipedia.org/wiki/Autoencoder>.
- Banaee, H, and A Loutfi. 2014. "Using conceptual spaces to model domain knowledge in data-to-text systems." *8th International Natural Language Generation (INLG) Conference, 19-21 June, Philadelphia, Pennsylvania, USA. Association for Computational Linguist*.
- Bao, J, M Dean, C Partridge, A Swami, and W J Leland. 2011. "Towards a theory of semantic communication." *In 2011 IEEE Network Science Workshop*. IEEE. 110-117.
- Barbosa, L S, W Marshall, S Streipert, L Albantakis, and G Tononi. 2020. "A measure for intrinsic information." *Scientific reports* 10(1), 1-9.
- Bar-Hillel, Y, and R Carnap. 1953. *Outline of a theory of semantic information*. Massachusetts Institute of Technology. Research Laboratory of Electronics.
- Bar-Yam, Y. 2004. "Multiscale variety in complex systems." *Complexity* 9(4) 37-45.
- Bateson, G. 1972. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.
- Battaglia, P W, J B Hamrick, V Bapst, A Sanchez-Gonzalez, V Zambaldi, M Malinowski, A Tacchetti, et al. 2018. "Relational inductive biases, deep learning, and graph networks." *arXiv preprint arXiv:1806.01261*.
- Beaumont, R. 2020. *Image embeddings*. July 20. <https://rom1504.medium.com/image-embeddings-ed1b194d113e>.
- Bedau, M A. 1997. "Weak emergence." *Philosophical Perspectives*. 11, 375–399.
- Blackmore, S. 2005. *Consciousness: A very short introduction*. Oxford University Press; 1st edition.
- Brouwer, E J. 1911. "Über Abbildungen von Mannigfaltigkeiten." *Math. Annalen* 71:97–115.
- n.d. *Buridan's ass*. https://en.wikipedia.org/wiki/Buridan%27s_ass.
- Carroll, S. 2016. *The big picture: On the origins of life, meaning, and the universe itself*. Dutton.
- Chaitin, G J. 1969. "On the length of programs for computing finite binary sequences: Statistical considerations." *Journal of the ACM* 16(1):145--159.
- Chaitin, G J. 2006. "The limits of reason." *Scientific American* 294(3), 74–81.

- Chalmers, D J. 1995. "Facing up to the problem of consciousness." *Journal of Consciousness Studies* 2 200-219.
- Chalmers, D J. 1992. "Subsymbolic computation and the chinese room." In *The symbolic and connectionist paradigms: Closing the gap*, by J. (Ed.) Dinsmore, 25-48. Psychology Press.
- . 1996. *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.
- Chomsky, N. 1990. "On the nature, acquisition and use of language." In *Mind and cognition: A reader*, by W.G. Lycan (ed.), 627-645. Cambridge MA and London UK: Blackwells.
- . 1957. *Syntactic structures*. Mouton.
- n.d. *Chunking_(psychology)*. [https://en.wikipedia.org/wiki/Chunking_\(psychology\)](https://en.wikipedia.org/wiki/Chunking_(psychology)).
- n.d. *Conway's Game of Life*. https://en.wikipedia.org/wiki/Conway's_Game_of_Life.
- Coradeschi, S, A Loutfi, and B Wrede. 2013. "A short review of symbol grounding in robotic and intelligent systems." *KI-Künstliche Intelligenz* 27.2: 129-136.
- Corning, P A. 2002. "The re-emergence of "emergence": A venerable concept in search of a theory." *Complexity*, 7 (6) 18-30.
- Crutchfield, J P, and K Young. 1990. "Computation at the onset of chaos." In *Entropy, complexity, and the physics of information. SFI studies in the sciences of complexity, VIII.*, by W Zurek, 223–269. Reading, Massachusetts: Addison-Wesley.
- Davidson, D. 1970. "Mental events." In *Essays on actions and events (Reprint)*. Oxford: Clarendon Press.
- Davidson, D. 1967. "Truth and meaning." *Synthese* 17 (3): 304–323.
- Davies, P. 2014. "Universe from bit." In *Information and the nature of reality: From physics to metaphysics.*, by P Davies and N H Gregersen. Cambridge University Press.
- Davies, P, and N H Gregersen. 2014. *Information and the nature of reality: From physics to metaphysics*. Cambridge University Press.
- Deledalle, G. 2000. *Charles S. Peirce's philosophy of signs - essays in comparative semiotics*. Bloomington: Indiana University Press.
- Devlin, J, M W Chang, K Lee, and K Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.
- n.d. *Digital_physics*. https://en.wikipedia.org/wiki/Digital_physics.
- Doerig, A, A Schurger, K Hess, and M H Herzog. 2019. "The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness." *Consciousness and cognition* 72, 49-59.
- Dretske, F I. 1981. *Knowledge and the flow of information*. Oxford: Blackwell; reprinted, Stanford, CA: CSLI Publications, 1999.
- Dreyfus, H L. 1972. *What computers can't do: a critique of artificial reason (1st edition)*. Harper & Row.

- Dreyfus, H L. 2007. "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian." *Philosophical psychology* 20.2, 247-268.
- Džeroski, S, L De Raedt, and K Driessens. 2001. "Relational reinforcement learning." *Machine learning* 43.1-2: 7-52.
- Fantino, E, and S Stolarz-Fantino. 2013. "The logic and illogic of human reasoning." In *APA handbooks in psychology. APA handbook of behavior analysis, Vol. 1. Methods and principles*, by G J Madden, W V Dube, T D Hackenberg, G P Hanley and K A Lattal , 439–461. American Psychological Association.
- Feinberg, T E, and J Mallatt. 2020. "Phenomenal consciousness and emergence: Eliminating the explanatory gap." *Frontiers in Psychology* 11 1041.
- Floridi, L. 2005. "Is information meaningful data?" *Philosophy and Phenomenological Research* 70(2): 351–370.
- Floridi, L. 2019. "Semantic conceptions of information." In *The Stanford Encyclopedia of Philosophy*,, by E N Zalta, <https://plato.stanford.edu/archives/win2019/entries/information-semantic/>.
- . 2011. *The philosophy of information*. Oxford University Press.
- n.d. *Four causes*. https://en.wikipedia.org/wiki/Four_causes.
- Freeman, W J. 1999. *How brains make up their minds*. Weidenfeld & Nicolson.
- Frege, G. 1892. "On sense and reference" ["Über sinn und bedeutung"]." *Zeitschrift für Philosophie und philosophische Kritik*, vol. 100.
- Friston, K. 2010. "The free-energy principle: A unified brain theory?" *Nature reviews neuroscience* 11.2: 127-138.
- Friston, K. 2012. "The history of the future of the Bayesian brain." *Neuroimage* 62(2):1230-1233.
- n.d. *F-score*. <https://en.wikipedia.org/wiki/F-score>.
- Gärdenfors, P. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, Massachusetts: MIT Press.
- Gibson, J J. 1979. *The ecological approach to visual perception*. Boston: Houghton Mifflin Harcourt.
- . 1966. *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
- Gleick, J. 2011. *The information: A history, a theory, a flood*. Random House.
- Goan, E, and C Fookes. 2020. "Bayesian neural networks: An introduction and survey." In *Case Studies in Applied Bayesian Data Science*., 45-87. Springer, Cham.
- Goddard, C, and A Wierzbicka. 2014. "Semantic fieldwork and lexical universals." *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 38.1: 80-127.

- Gödel, K. 1934. "On undecidable propositions of formal mathematical systems." (*mimeographed lecture notes; taken by S. Kleene and J. Rosser*), reprinted with corrections in Davis 1965, 41–81, and Gödel 1986, 346–371.
- Goertzel, B. 2015. "Artificial general intelligence." *Scholarpedia* 10(11):31847.
- Goldman, A I. 2012. *Theory of mind*. The Oxford handbook of philosophy of cognitive science.
- Goldstein, J. 1999. "Emergence as a construct: History and issues." *Emergence* 11: 49-72.
- Green, M, and V Evans. 2006. "Cognitive linguistics: An introduction." In *Cognitive Linguistics*, by V Evans. Edinburgh.
- Grice, H P. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hacohen, G, and D Weinshall. 2019. "On the power of curriculum learning in training deep networks." *arXiv preprint arXiv:1904.03626*.
- Haken, H, and J Portugali. 2014. *Information adaptation: The interplay between shannon information and semantic information in cognition*. Springer.
- Harnad, S. 1990. "Against computational hermeneutics." *Social Epistemology* 4 (Social Epistemology) 167-172.
- Harnad, S. 1987. "Category induction and representation." In *Categorical perception: The groundwork of cognition*, by S. Harnad (Ed.). New York: Cambridge University Press.
- Harnad, S. 2007. "Symbol grounding problem." *Scholarpedia*, 2(7):2373.
- Harnad, S. 1990. "The symbol grounding problem." *Physica D* 42 (Physica D 42: 335-346.) 335-346.
- Hayes, P. 1979. "The naive physics manifesto." In *Expert Systems in the micro-electronic age*, by D Michie. Edinburgh: Edinburgh University Press.
- Hintikka, J, and P Suppes. 1970. *Information and inference*. Dordrecht: Reidel.
- Hochreiter, S, and J Schmidhuber. 1997. "Long short-term memory." *Neural computation* 9.8: 1735-1780.
- Hoel, E P. 2018. "Agent above, atom below: How agents causally emerge from their underlying microphysics." In *In wandering towards a goal*, 63-76. Springer, Cham.
- Hoffman, D D. 2019. *The case against reality: How evolution hid the truth from our eyes*. Allen Lane.
- Hofstadter, D, and E Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.
- Holland, J H. 1995. *Hidden order: How adaptation builds complexity*. New York, NY: Addison-Wesley.
- n.d. *Homeostat*. <https://en.wikipedia.org/wiki/Homeostat>.
- Hopfield, J J. 1982. "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the National Academy of Sciences of the USA* 79: 8 2554–2558.

- Howard, R A. 1966. "Information value theory." *IEEE Transactions on systems science and cybernetics* 2.1 22-26.
- n.d. *Hubert Dreyfus's views on artificial intelligence*.
https://en.wikipedia.org/wiki/Hubert_Dreyfus%27s_views_on_artificial_intelligence.
- Hutter, M. 2007. "Algorithmic information theory." *Scholarpedia*, 2(3):2519 2(3):2519.
- . 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer, Berlin.
- Jiang, J, D Chen, T Huang, and Z Lu. 2018. "Graph convolutional reinforcement learning." *arXiv preprint arXiv:1810.09202*.
- Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kastrup, B. 2019. *The idea of the world: A multi-disciplinary argument for the mental nature of reality*. John Hunt Publishing.
- n.d. *Knowledge Representation and Reasoning*.
https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning.
- n.d. *Knowledge_representation_and_reasoning*.
https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning.
- Kolchinsky, A, and D H Wolpert. 2018. "Semantic information, autonomous agency and non-equilibrium statistical physics." *Interface Focus* 8.6: 20180041.
- Kolmogorov, A. 1963. "On tables of random numbers." *Sankhyā Ser. A*. 25: 369–375.
- Korzybski, A. 1933. "Science and sanity. An introduction to non-Aristotelian systems and general semantics." *The International Non-Aristotelian Library Pub. Co.* 747–761.
- Kotseruba, I, and J K Tsotsos. 2020. "40 years of cognitive architectures: Core cognitive abilities and practical applications." *Artificial Intelligence Review* 53.1: 17-94.
- Krizhevsky, A, I Sutskever, and G E Hinton. 2017. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60.6: 84-90.
- Kullback, S, and R A Leibler. 1951. "On information and sufficiency." *Annals of Mathematical Statistics* 22 (1): 79–86.
- Lake, B M, T F Ullman, T D Tenenbaum, and S J Gershman. 2017. "Building machines that learn and think like people." *Behavioral and brain sciences* 40.
- Lakoff, G, and M Johnson. 2003. *Metaphors we live by*. Chicago: University Of Chicago Press.
- Landauer, R. 1991. "Information is physical." *Physics Today* 44.5: 23-29.
- Langton, C G. 1990. "Computation at the edge of chaos." *Physica D*. 42: 12.
- n.d. *Laplace's_demon*. https://en.wikipedia.org/wiki/Laplace%27s_demon.

- Laughlin, S. 1981. "A simple coding procedure enhances a neuron's information capacity." *Zeitschrift für Naturforschung* c 36.9-10: 910-912.
- LeCun, Y, Y Bengio, and G E Hinton. 2015. "Deep learning." *Nature* 521(7553), 436-444.
- Lenat, D B. May, 1988. "The case for inelegance." *Proceedings of the International Workshop on Artificial Intelligence for Industrial Applications*. Tokyo.
- Lenat, D B, R V Guha, K Pittman, D Pratt, and M Shepherd. 1990. "Cyc: Toward programs with common sense." *Commun. ACM* 33 (8): 30–49.
- Levine, J. 1983. "Materialism and qualia: The explanatory gap." *Pacific Philosophical Quarterly*, 64 354-361.
- Lewis, C I. 1929. *Mind and The World Order: an Outline of a Theory of Knowledge*. New York: Charles Scribner's Sons.
- Lexico. n.d. "Cognition." Oxford University Press and Dictionary.com.
- Libet, B, C A Gleason, E W Wright, and D K Pearl. 1983. "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) – the unconscious initiation of a freely voluntary act." *Brain* 106 (3): 6.
- n.d. *List of cognitive biases*. https://en.wikipedia.org/wiki/List_of_cognitive_biases.
- Lloyd, S. 2006. *Programming the universe: a quantum computer scientist takes on the cosmos*. Vintage.
- Lopez-Suarez, M, I Neri, and L Gammaito. 2016. "Sub-k BT micro-electromechanical irreversible logic gate." *Nature communications* 7.1: 1-6.
- Lundgren, B. 2019. "Does semantic information need to be truthful?" *Synthese* 196.7: 2885-2906.
- Lynch, C, M Khansari, T Xiao, V Kumar, J Tompson, S Levine, and P Sermanet. 2020. "Learning latent plans from play." *In Conference on Robot Learning*. PMLR 1113-1132.
- Mallery, J C. 1988. "Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers." *Master's thesis, M.I.T. Political Science Department*.
- Matthias, A. 2019. *Common sense computing*. Moral Robots. January 24.
<https://medium.com/datadriveninvestor/cyc-making-computers-with-common-sense-daad1a27781c>.
- McClelland, J L, and D E Rumelhart. 1986. "The PDP research group." *Parallel distributed processing* 1.2.
- Melamed, Y Y, and M Lin. 2020. "Principle of sufficient reason." In *The Stanford encyclopedia of philosophy (Spring 2020 Edition)*, by E N Zalta,
<https://plato.stanford.edu/archives/spr2020/entries/sufficient-reason/>. Metaphysics Research Lab, Stanford University.
- Mercier, H, and D Sperber. 2017. *The enigma of reason*. Harvard University Press.

- Mikolov, T, K Chen, G Corrado, and J Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv:1301.3781*.
- Miller, G A, and E A Friedman. 1957. "The reconstruction of mutilated English texts." *Information and Control* 1.1: 38-55.
- Miller, J G. 1978. *Living Systems*. McGraw-Hill. <https://archive.org/details/LivingSystems>.
- n.d. *Minimum_description_length*. https://en.wikipedia.org/wiki/Minimum_description_length.
- n.d. *Mutual_assured_destruction*. https://en.wikipedia.org/wiki/Mutual_assured_destruction.
- Nassar, M R, C Helmers, and M J Frank. 2018. "Chunking as a rational strategy for lossy data compression in visual working memory." *Psychological review* 125.4 486.
- n.d. *Necker_cube*. https://en.wikipedia.org/wiki/Necker_cube.
- Newell, A, and H A Simon. 1976. "Computer science as empirical inquiry: Symbols and search." *Communications of the ACM* 19 (3): 113–126.
- n.d. *Ontology (information science)*. [https://en.wikipedia.org/wiki/Ontology_\(information_science\)](https://en.wikipedia.org/wiki/Ontology_(information_science)).
- Packard, N, J Crutchfield, D Farmer, and R Shaw. 1980. "Geometry from a time series." *Physical Review Letters* 45 (9): 712–716.
- Page, S E. 2009. *Understanding complexity - The Great Courses, 1st edition*. The Teaching Company.
- Pascual-García, A. 2018. "A constructive approach to the epistemological problem of emergence in complex systems." *PloS one* 13(10), e0206489.
- Pearl, J, and D Mackenzie. 2018. *The book of why: The new science of cause and effect*. Basic Books.
- Pickering, A. 2010. *The cybernetic brain: Sketches of another future*. University of Chicago Press.
- Pitts, M A, A Martínez, C Stalmaster, J L Nerger, and S A Hillyard. 2009. "Neural generators of ERPs linked with Necker cube reversals." *Psychophysiology* 46.4: 694-702.
- n.d. *Platt Labs*. <http://plattlabs.rocks/>.
- Prakash, C, K D Stephens, D D Hoffman, M Singh, and C Fields. 2020. "Fitness beats truth in the evolution of perception." *Acta Biotheoretica* 1-23.
- Prigogine, I, and I Stengers. 1997. *The end of certainty: Time, chaos and the new laws of nature*. The Free Press.
- Princeton University. 2010. *About WordNet*. WordNet. Princeton University. <https://wordnet.princeton.edu/>.
- n.d. *Principle of sufficient reason*. https://en.wikipedia.org/wiki/Principle_of_sufficient_reason#Leibniz's_view.
- Purves, D, and R B Lotto. 2003. *Why we see what we do: An empirical theory of vision*. Sinauer Associates.

- Ramsey, W. 2016. "Eliminative Materialism." In *The Stanford Encyclopedia of Philosophy (Winter 2016 ed.)*, by Edward N. (ed.). Zalta. Metaphysics Research Lab, Stanford University.
- Ren, P, Y Xiao, X Chang, P Y Huang, Z Li, X Chen, and X Wang. 2020. "A survey of deep active learning." *arXiv preprint arXiv:2009.00236*.
- Rorty, R. 1979. *Philosophy and the mirror of nature*. Princeton, NJ: Princeton University Press.
- Rosch, E, and B L Lloyd. 1978. *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum.
- Rosen, R. 2012. *Anticipatory systems: Philosophical, mathematical, and methodological foundations (2nd edition)*. New York: Springer.
- Rovelli, C. 2015. "Relative information at the foundation of physics." In *It from bit or bit from it?*, by A Aguirre, B Foster and Z Merali, 79-86. Springer, Cham.
- Rowlands, M. 2010. *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.
- n.d. *Rubin_vase*. https://en.wikipedia.org/wiki/Rubin_vase.
- Ruble, D N, L S Newman, W S Rholes, and J Altshuler. 1988. "Children's "naïve psychology": The use of behavioral and situational information for the prediction of behavior." *Cognitive Development* 3(1), 89-112.
- Ryle, G. 1949. *The concept of mind*. Chicago: University of Chicago Press.
- Schaal, S. 1997. "Learning from demonstration." *Advances in neural information processing systems* 1040-1046.
- Searle, J. 1980. "Minds, brains and programs." *Behavioral and Brain Sciences* 3 (3): 417–457.
- Settles, B. 2010. "Active learning literature survey." *Computer Sciences Technical Report 1648. University of Wisconsin–Madison*.
- Shanahan, M. 2016. "The frame problem." In *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)*, by E N Zalta, URL = <<https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>>.
- Shannon, C E. 1948. "A mathematical theory of communication." *The Bell system technical journal* (The Bell system technical journal 27 (3), 379-423) 27 (3), 379-423.
- Shannon, C E. 1951. "Prediction and entropy of printed English." *Bell System Tech. J.* 30, 50-64.
- Siegelmann, H T. 2003. "Neural and super-Turing computing." *Minds and Machines* 13.1: 103-114.
- Siewert, C. 2017. "Consciousness and intentionality." In *The Stanford Encyclopedia of Philosophy*, by E N Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2017/entries/consciousness-intentionality>.

- Silver, D, T Hubert, J Schrittwieser, I Antonoglou, M Lai, A Guez, M Lanctot, et al. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science*, 362(6419) 1140-1144.
- Smolensky, P. 1988. "On the proper treatment of connectionism." *Behavioral and brain sciences* 11.1: 1-23.
- Smolensky, P. 2012. "Subsymbolic computation theory for the human intuitive processor." *Conference on Computability in Europe. Springer, Berlin, Heidelberg*.
- Solomonoff, R. 1964a. "A formal theory of inductive inference Part I." *Information and Control* 7 (1): 1–22.
- Solomonoff, R. 1964b. "A formal theory of inductive inference Part II." *Information and Control* 7 (2): 224–254.
- Speer, R, J Chin, and C Havasi. 2017. "ConceptNet 5.5: An open multilingual graph of general knowledge." *Proceedings of AAAI* 31.
- Spirtes, P, and K Zhang. 2016. "Causal discovery and inference: Concepts and recent methodological advances." *In Applied informatics* 3(1): 1-28). SpringerOpen.
- Steels, L. 2008. "The symbol grounding problem has been solved. So what's next." *In Symbols and embodiment: Debates on meaning and cognition*, 223-244.
- Sterling, P. 2012. "Allostasis: A model of predictive regulation." *Physiology & Behavior* 106 (1): 5–15.
- Takens, F. 1981. "Detecting strange attractors in turbulence." *In Dynamical systems and turbulence, lecture notes in mathematics, vol. 898*, by D A Rand and L -S Young, 366–381. Springer-Verlag.
- Talmy, L. 2000. "Force dynamics in language and cognition." *In Toward a cognitive semantics vol I: Concept structuring systems.*, by L Talmy. Cambridge: MIT Press.
- Tarski, A. 1935. *The Concept of truth in formalized languages. Logic, semantics, metamathematics*. Indianapolis: Hackett.
- Tegmark, M. 2017. *Life 3.0: Being human in the age of artificial intelligence (1st edition)*. New York: Knopf.
- n.d. *Theory_of_forms*. https://en.wikipedia.org/wiki/Theory_of_forms.
- Todorovic, D. 2008. "Gestalt principles." *Scholarpedia* 3(12):5345.
- Tononi, G. 2015. "Integrated information theory." *Scholarpedia* 10(1):4164.
- Tononi, G, M Boly, M Massimini, and C Koch. 2016. "Integrated information theory: From consciousness to its physical substrate." *Nature Reviews Neuroscience* 17 (7): 450–461.
- Tum, P. 2020. "A survey of the state-of-the-art language models up to early 2020." *Medium*. February 22. <https://medium.com/@phylypo/a-survey-of-the-state-of-the-art-language-models-up-to-early-2020-aba824302c6>.

- Vaswani, A, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 5998-6008.
- Weaver, W. 1953. "Recent contributions to the mathematical theory of communication." *ETC: A review of general semantics* 261-281.
- Weinberger, E D. 2002. "A theory of pragmatic information and its application to the quasi-species model of biological evolution." *Biosystems* 66.3: 105-119.
- Whitehead, A N, D R Griffin, and D W Sherburne. 1979. *Process and reality: An essay in cosmology*. New York: Free Press.
- Wierzbicka, A. 1996. *Semantics: Primes and universals*. Oxford University Press.
- Wigner, E P. 1961. "Remarks on the mind-body question." In *The scientist speculates*, by I J Good. London: Heinemann.
- Wigner, E P. 1960. "The unreasonable effectiveness of mathematics in the natural sciences." *Communications on Pure and Applied Mathematics*.
- Wittgenstein, L. 1922. *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul.
- Wolfram, S. 2002. *A new kind of science*. Champaign, IL: Wolfram Media.
- Yan, S, S Zhang, and X He. 2019. "A dual attention network with semantic embedding for few-shot learning." *AAAI* 9079-9086.
- Zambaldi, V, D Raposo, A Santoro, V Bapst, Y Li, I Babuschkin, K Tuyls, et al. 2018. "Relational deep reinforcement learning." *arXiv preprint arXiv:1806.01830*.
- n.d. *Zeno's paradoxes*. https://en.wikipedia.org/wiki/Zeno%27s_paradoxes#Dichotomy_paradox.
- Zimmermann, J. 2015. *Hermeneutics: A very short introduction*. OUP Oxford.
- n.d. *Zone of proximal development*. https://en.wikipedia.org/wiki/Zone_of_proximal_development.