An Analysis and Proposed Design for Artificial General Intelligence

1st Brian Tang
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA
bjaytang@umich.edu

Abstract

In this work, we present context for the development of an artificial general intelligence (AGI) and discuss shortcomings and strengths of current approaches in AGI. We propose a preliminary design and evaluation methodology for AGI, and describe several different embodiment types for our design. Finally, we discuss future prospects and timelines for AGI.

Index Terms

Artificial general intelligence, artificial intelligence, machine learning, deep learning, cognition, intelligence, cognitive science

I. Introduction

Today, we have many different specialized artificial intelligence (AI), typically created using machine learning (ML) and empirical risk minimization (ERM), which are able to achieve a human-like level of performance on specific tasks. However, these narrow forms of AI are unable to handle complex tasks and general understanding of the environment around them. Artificial general intelligence (AGI) is one of the biggest challenges in the field of AI. How can AGI achieve a human-level (or even just a more robust level) of cognitive ability? In this paper, we outline the defining characteristics of AGIs, several cognitive capabilities and structures that an AGI can leverage when interacting with its environment, and a potential design for connecting each cognitive capability within an AGI system.

AGI need not necessarily be infinitely generalizable. Rather, creating an AGI with the ability to generalize its knowledge and handle a variety of goals and tasks in different environments would be a challenging enough task on its own.

Intelligence can be defined as how well non-learning agents are able to use the knowledge available to perform a task. While this definition seemingly excludes learning as a measure of intelligence, the task or goal of the agent could be many different concepts. The set of possible tasks could include perception, generating new knowledge, reasoning, learning, self-motivation, perception, or even metacognition.

II. Background

Researchers in AGI have taken several approaches to constructing and designing AGI systems. The symbolic approach uses a centralized control of perception, cognition, and action. While symbolic processing allows for broad generalizations, most architectures are incapable of giving rise to complex structures and dynamics. Several examples include Markov logic networks [1], inductive logic programming, ACT-R [2], Cyc [3], and SOAR [4]. The emergentist approach expects abstract symbolic processing to emerge from low-level data processing dynamics. Though this approach has shown to be useful for patterns and associative memory, it may struggle to automatically organize its own architectures. Among these approaches are deep learning, reinforcement learning, and pattern recognition systems. The universalist approach suggests that a meta-algorithm exists to automatically improve and optimize an AGI's intelligence. While this approach is more rigorous and ideal, it is computationally infeasible. There are other fields as well, such as computational neuroscience and developmental robotics. These approaches investigate different ways to model the brain and program robotics to learn using intrinsic motivation. Finally, the hybrid approach argues that a combination of the above approaches is needed. While several systems like CLARION [5] and DUAL use this approach, if any underlying components or systems are inadequate, this could result in brittleness or weak robustness. For our AGI design, we leverage the advances in both explicit symbol processing and high-dimensional statistical learning in order to create a more practical method of achieving human-like cognitive capabilities.

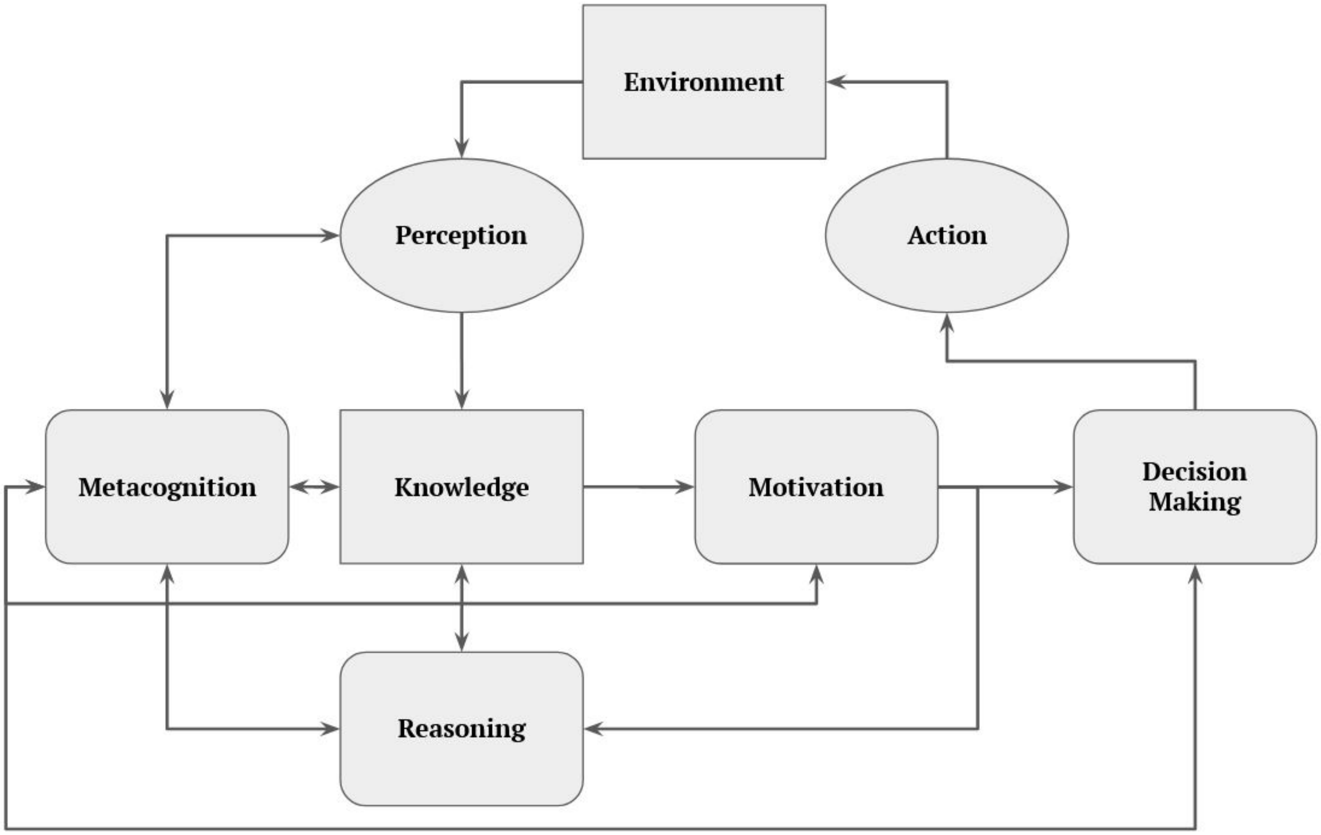

Fig. 1. An high level overview of the different components of the AGI design.

A. Scales of Intelligence

1) Evolution of Intelligence: There is ladder of intelligence similar to Maslow's Hierarchy of Needs [6]. It is likely that the higher rungs of reasoning rely on an agent having all the prior capacities provided in the lower rungs of reasoning [7]. Thus, creating a human-level AGI will likely rely on a "bottom-up" approach for these cognitive capabilities. In Section III, we describe how each component of our design enables each of the rungs of reasoning.

- **[Asocial Reasoning]:** Capability of hiding, foraging, hunting, killing, fleeing, eating, figuring out what things are and where things are. These asocial reasoning capabilities can be reduced to the cycle of perceiving, reasoning, deciding, and acting.
- **[Social Reasoning]:** Capability of nurturing, protecting, feeding, bonding, giving, sharing with others. These social reasoning capabilities can be reduced to an integration of basic social and ethics concepts within an agent's motivational system.
- **[Animal Cultural Reasoning]:** Capability of following, cooperating, playing, leading, warning, tricking, stealing, teaching. These capabilities are similar to the social reasoning rung.
- **[Oral Linguistic Reasoning]:** Capability of promising, apologizing, giving oaths, agreeing, making covenants, naming things. These capabilities build on the previous rungs by adding a layer of abstraction and integrating a complex understanding of time, space, and context into social reasoning.
- **[Literate Reasoning]:** Concepts like libraries, writing, contracts, fiction, technology, and documentation. These concepts emerge from a society of agents capable of communicating, understanding, and documenting abstract concepts using imagery, language, etc.
- **[Civilization-scale Reasoning]:** Concepts like democracy, empires, philosophy, science, mathematics, culture, economy, nations, and literature. These civilization-level concepts arise from having a large population of agents with large amounts of computational power to allocate towards these higher rung reasoning abilities (rather than survival).

2) Animal Intelligence: It is unclear exactly which cognitive capabilities are unique to humans vs. animals. While animals may be capable of understanding semantics and syntactic structure, they may be incapable of *generating* complex grammars and syntactic structures [8]. Animals also struggle to reason with numbers, symbols, logic, and language well. However, they are able to perceive objects, model physical properties, and reason about them. While this is also difficult for young human children, over time, they are able to learn and reason about these concepts. While humans and animals share the same underlying components and structures (e.g., neurons, brain sections), there are sets of cognitive capabilities which are unavailable to animals. While animals may have larger overall brain size, the proportion of the human brain relative to body size is remarkably larger, especially in the case of the human anterior cortex being orders of magnitude larger.

3) Ingredients of Human Intelligence: Kinzler *et al.*present a potential set of core ingredients for human intelligence [9]. The first of which includes an understanding of numbers, space, physics, and psychology For example, building an understanding of verbal and nonverbal cues for attention and distinguishing agents from objects would be a helpful prior to encode in an AGI. As agents are expected to have goals, take actions, and reason about the world, it would be practical to represent them differently from other concepts such as objects, places, or time. Rapid model building can also enable more complex representations, and thus reasoning. For example, human adults and children especially excel at few-shot learning of concepts before humans can understand things. Lake *et al.*argue that compositionality, causality, and meta learning allow humans to build (causal) models [10]. These capabilities appear to support the interactions underlying and connecting different concepts together. These capabilities also allow for generalizing and transferring knowledge acquired in a particular domain to other similar environments. Finally, the ability to think quickly using approximation, extrapolation, and simple, associative model-free learning allow humans to function and meet fast and strict real-time constraints.

4) Time Scales of Human Action: Newell *et al.*derive a set of time scales at which different human cognitive processes function [11]. Human cognitive architecture is a hierarchy of system levels. At different levels of abstraction and computation speed are collections of components that are linked together and interact. In both humans and computers, there are real-time constraints on cognition which must be met. For example, sentences may be uttered in 1 second, and an agent's reaction to it should take around 1 second to process. Intelligence is typically most evident when an agent is observed over longer time scales (rational and social bands). Note that this does not imply that the biological and cognitive band of time scales are not useful. Cognitive processes that function at these smaller time scales are important for the underlying functioning of the agent, as well as providing robustness to the cognition of an agent. Ideally, an AGI should function at all of the following time scales:

- **[Biological Band]:** This time scale is within the range of microseconds to milliseconds. In humans, these include the operations and activations of organelles, neurons, and neural circuitry.
- **[Cognitive Band]:** This time scale is within the range of milliseconds to seconds. In humans, these include unit tasks, operations, and deliberate acts.
- **[Rational Band]:** This time scale is within the range of minutes to hours. In humans, these take the form of tasks.
- **[Social Band]:** This time scale is within the range of days to months.
- **[Historical Band]:** This time scale is within the range of years to millennia.
- **[Evolutionary Band]:** This time scale is within the range of millions of years.

We can further divide the cognitive and rational time scales into a finer granularity.

- **[100 Milliseconds]:** This includes reactionary decisions, skilled (prepared) behavior, primitive internal actions, and access to long-term memories.
- **[1 Second]:** This includes simple reasoning, mental imagery access, and language processing.
- **[10 Seconds]:** This includes complex reasoning, analogy, planning, meta reasoning, and theory of mind.

III. Human-level AGI Design

This paper will be organized into the most crucial components of AGI.

Our design assumes that an AGI can be created using a centralized architecture containing metacognitive, motivation, and knowledge components. Similar to many of the cognitive architectures encapsulated by the common model of cognition [12], each component is interconnected and has its own dedicated processes which manage the allocation of computation and memory resources. The AGI system contains many sub-components which can run in parallel including learning, perception, reasoning, memory retrieval, emotion appraisal, representational modeling, and more. Our AGI has been designed to provide the underlying functions that enable the many of the capabilities general intelligence provides (Section XIII-B) and reasoning at each time scale (Section II-A4).

We plan to leverage many of the advances in deep learning [13] (for perception, representations, language) along with the capabilities of a symbolic reasoning system (for performing abduction and model building on representations to learn complex interactions in physics, psychology, etc.). The neural network architectures would excel at covering the lower bands of Newell's levels, while the symbol system could be used to represent the higher level abstractions and cognitive capabilities. By designing the knowledge base to represent information at different levels of abstraction, this would allow the agent to engage in world/model building at different granularities. For example, model building could be as vague as constructing a scene of objects in a setting and their interactions. Or, it could also be fine-grained with building conceptual models of physics or mechanisms within objects.

We opt to create a dynamic cognitive architecture in order to encode architectures themselves as "learnable knowledge". For example, to improve object recognition/categorization at a meta-level, the agent can initialize a DNN to a set of well-performing weights. Afterwards, this pre-training procedure could be represented in knowledge. Compared to a static architecture, a dynamic architecture allows for increased modularity and flexibility. This approach does incur the drawback of reduced efficiency and supervision.

Our AGI design will not learn fundamental concepts and mechanisms from scratch. These include perceptual features, attention mechanisms, natural language, causal learning, and more. Since we have opted for a hybrid design, there is little reason not to leverage and adjust existing pretrained ML models designed to perform specialized tasks.

IV. Environment

We argue that an environment must be sufficiently complex for generally intelligent agents to demonstrate intelligent behavior. Complex intelligent capabilities cannot emerge from AGI which learn from simple or constrained environments. As such, several different environments are candidates that can be rich enough to support AGI developments. Some examples include real-world embodiment, the internet, simulated environments that emulate the real world, and complex multi-modal datasets (e.g., self-driving datasets with plentiful sensor information).

A. Real-World Embodiment

It is crucial that human-like AGI are embodied in real-world environments. This not only makes interaction with the environment easier for the agent, but it also grounds the agent's understanding in the same world we live in. This common grounding experience can reduce disconnects between the understanding of concepts experienced by both humans and AGI systems.

B. Curriculum

Common-sense knowledge and reasoning can be particularly useful for teaching an AGI to map language to concepts. It can also allow for a system to categorize and add properties to its objects, agents, and settings. In doing this, the AGI can create more generalizations after having some modal data surrounding common-sense and common-ground information.

V. Perception and Actuation

A. Environment Sensors

A human-like AGI should be able to sense information from its environment through various modalities. The most crucial sensor capabilities would include a real-time camera and audio feed. Including proximity sensors, GPS and accelerometers could improve the richness of information for navigation and motion.

B. Feature Bias

Using pretrained models for image, audio, and language processing, we can extract the core features that deep neural networks (DNNs) bias their decisions towards. These include color, edges, shapes, motion, frequency, amplitude, grammar, entities, word associations, etc. Using DNNs to extract and represent visual/auditory concepts would allow symbols to ground their representations in real-world data.

C. Attention

Whenever attention resources in humans are stretched, the finer distinctions fail to guide object representations. In the case of perception, we may fail to focus on the fine-grained details within perceived objects, agents, and settings. When performing memory retrieval, fewer properties may be extracted due to time limitations. Attention serves as a mechanism to access different abstraction levels from represented knowledge. It also helps drastically reduce the amount of information required for processing. While we can hard-code some attentional priors (such as through surprise, direct references to the agent), attention can also be intentionally guided by the internal goals and motivations of the AGI.

D. Actuators

An AGI must interact with its environment to some degree or at least output its observations. Any AGI implemented using our design should have some amount of agency/autonomy over itself and its environment in order to close the perception → decision → action → evaluation loop. Human-like AGIs could communicate using speakers and digital displays (for text and emotion), and move/interact through its environment with motors, wheels, arms, etc.

VI. Knowledge

A. Core Systems and Concepts

Concepts are abstract and can have visual, audio, and language representations that can be constructed from perceived data or by imagining combinations of properties or core systems. To better categorize concepts, we divide them into three core systems. In this way, concepts can be constructed in our AGI system using innate sets of properties and interactions. We define three core systems: objects, agents, and settings. Each core system has its own universal set of properties and relations with other core systems. It is important to note that a concept could be found in more than one core system category depending on the context. For example, a car could be seen as an object (physical properties), agent (someone driving the car and using it as a medium for action), or setting (one can perceive the car as the environment in which other events occur).

1) Objects: Objects represent any non-agent entity that is present in the environment. Several properties of objects include numbers, color, shape, usages, sub-objects, etc.

2) Agents: Agents represent any entity capable of having goals, making decisions to achieve those goals, and performing actions (that impact the environment) which are in line with their decisions. Agents are more costly concepts, as they are simplified projections of internal understandings of cognitive systems onto entities in the environment (Theory of Mind). Like an AGI, agents have their own sets of capabilities, goals, and actions.

3) Settings: Settings represent the time and space (place) in which a scene of episodic memory takes place. Settings can also be represented as a standalone semantic concept.

4) Actions: Actions are concepts that are unique in that they describe relations between other concepts. Executions and actions can be used as symbols/nodes as well. Actions and skills can comprise of sub-actions as well as the typical representational data forms.

5) Interactions: Whenever a concept is subsumed by another concept, either 1) a new concept may be formed, or 2) the subsumed concept becomes a property of the "engulfing" concept.

Whenever concepts interact with each other, the interaction itself could be created as 1) a new concept, or 2) a property of the "edge" linking the concepts.

Objects can interact with other objects through physical laws of motion, momentum, etc. An object may also have many smaller object concepts embedded within it to many degrees of abstraction (the atoms, inside the molecules, inside the fuel, inside the booster of a rocket).

Agents can interact with objects by manipulating or controlling them. In example of a person driving a car, an object envelops the agent and either a car with a person driving it is created, or the person becomes a property of the car. If a person were to use a hammer as a tool, the agent would become a person with a hammer as either a new concept or new property.

Finally, objects and agents are usually embedded within a larger setting. This setting is important for context for a particular scene, and it also helps construct the semantic concept of a setting. For example, a hospital over time can become associated with the objects and agents within it (doctors, patients, beds, medicine, MRI machines, etc.).

B. Sensory Representations

Concepts are represented in episodic memory as scenes which contain sets of objects, agents, and a setting. They can also be represented by symbols for quick operations which do not rely on visual or spatial representations (Section VI-C). These concepts are grounded in real-world sensory experiences and can also be represented in semantic memory as combination of properties. In semantic representations of concepts, the AGI system will store different granularity levels of details. For fast retrieval, the system will use the simplified concept which contains only the most essential underlying attributes and associations. For abstract concepts with no grounding in sensory representations, the sensory representations can be generated or extracted using neighboring concepts.

C. Symbols

Symbols are tokens which represent concepts. They are used solely as identifiers and operations for reasoning. The symbols can be mapped onto natural language relatively easily. Using natural language models such as GPT-3 or PaLM can allow the AGI to map stored knowledge into communicable symbols. For example, DALL-E has seen success in image-text related tasks. Developing models which work with multimodal representations of a symbol could enable more generalizable procedurally generated knowledge.

1) Abstraction: In order to avoid the Homunculus Fallacy [14] or infinitely recurring abstractions, our AGI system can define recursion limits in both creation of new concepts/properties/links and retrievals. These limits will be determined by a heuristic combining real-time constraints, data efficiency, and the usefulness of the abstraction.

We provide an example case study for recognition and categorization. Where DNNs perform pattern matching and associative learning for a particular set of classes and data inputs, our goal is to allow for pattern matching at every granularity. We can represent episodic memory as scenes with differing granularities. For example, an image of a driving scene can have objects arranged in a setting, and these might be matched with existing knowledge. Each object within a driving scene has their own sets of features (shapes, colors, motions, etc.) which can also be matched with knowledge. These objects may have subparts such as mechanical functions, chemical makeup, or physics that are impossible to directly perceive. However, these subpart scenes may not be relevant to our AGI. This abstraction will be applied to other types of concepts too such as those which represent decisions, plans, reasoning, actions, etc.

2) Language and Communication: While certain symbols could be ambiguous, for the most part, words and phrases provide a simple mechanism for mapping language to symbols.

D. Memory

Knowledge is stored in a graph with each node corresponding to concepts and their relevant episodic and semantic representations.

1) Short-term and Long-term Memory: Short-term memory contains the ability to hold onto recent states of the environment. After a period of downtime, short-term memory is transferred to long-term memory by converting state knowledge into episodic and semantic memory. This effectively results in a compression of memory and experiences without losing the essentials of the short-term memory.

2) Episodic Memory: Episodic memory contains entire scenes of experiences, each complete with objects, agents, and settings. These provide context for concepts (e.g., usage, relations, etc.). The AGI can use a composition of symbols, each with representations grounded in different sensor modalities to represent episodic memory. For example, it can construct a single scene with setting, objects, agents, and actions, and it can save snapshots of these scenes, with emphasis on recalling scenes with major differences.

3) Semantic Memory: Semantic memory contains the essential features and properties of a concept (e.g., shape, color, etc.).

E. Retrieval

Retrieval of knowledge is likely the biggest bottleneck to real-time performance of our AGI memory system. Due to the infinite depth of abstractions in our concept design, retrieval must be limited based on the amount of time available. Alternatively, we can construct a graph of neighboring concepts which are close together in some embedding space. This categorization reduces the amount of entities needed to search. Heuristics such as features, labels, heuristics can be computed using classification or perceptual similarity. These heuristics can provide "shortcut" connections between concepts. These connections can improve both retrieval speed and effectiveness. Retrieval can work in both a bottom-up or top-down fashion. E.g., seeing a cow's features, behaviors, senses, and other associations for the cow can activate the symbol of a cow. Alternatively, thinking of the symbol of a cow can anticipate the features, behaviors, senses, and other associations for the cow. This anticipatory inference can be combined with perception and reward learning for a predictive inference mechanism that improves retrieval over time (anticipation → perception → negative/positive reinforcement). This inference can rely on several factors such as perceptual similarity, frequency, and recency of retrieval. One of the functions of the metacognitive unit improves upon the actual information stored in association with that concept by creating a new concept with different properties/links/granularity.

1) Analogies: At higher levels of abstraction and coarse-granularity, concepts will begin to overlap and share many properties. Concepts with similar physical or causal properties begin to cluster. Analogies can act as a shortcut for creating properties for new concepts by directly mapping existing concepts. They can also reduce retrieval time. Properties of base representations can be mapped onto target representations both to understand analogies communicated to the AGI and to generate new connections between concepts. Similar to abstraction, analogies can be evaluated using correctness, robustness to concept representations, data efficiency, computational complexity, and usefulness.

2) Search: Knowledge both controls and informs search. There are fast and fixed mechanisms for searching over existing organized knowledge. In this case, there is no awareness of processing (Kahneman's System 1). However, problem search is serial, generative, and combinatorial. While it is also controlled by knowledge, there is awareness of processing (Kahneman's System 2). Note that both knowledge search and problem search can improve with experience.

Converting problem search into knowledge search can significantly speed up runtime. However, this typically requires a large amount of preparation and intuition into the context. For example, previous studies around human performance time and the Seibel task and cigar making have suggested that improvement in human cognition speed for practiced tasks seemingly resemble a power law scaling [15]. However, there is a notable trade-off in time spent preparing vs. time spend analyzing and deliberating.

We can compute indexes for recurring patterns to improve search speed. These indexes could map to existing episodic or semantic memories. In this manner, "common-sense" knowledge is the set of knowledge which an agent is exposed to frequently enough such that retrieval becomes a quick knowledge search.

VII. Reasoning

A. Symbolic Reasoning

Languages of thought could provide the ability to capture logical operations using properties and relations to create new states. The operational "functions" could be extracted from the analogy and abstraction mechanisms. This system would also allow for quantitative and numeric reasoning. Whenever there is downtime, the agent can procedurally imagine scenarios, plan, learn from existing knowledge. Otherwise, it may make inferences when it is relevant or a necessary requisite to achieving some goal. It limits this scope in accordance to the current/relevant goals of the agent. Finally, using induction (extrapolations of past experiences), deduction (using logical/deductive arguments), and abduction (composing hypothetical combinations of symbols/concepts), the agent generates beliefs and new scenarios. The AGI must combine new sources of knowledge with previous ones and updating the symbol, until it becomes some average or aggregate of experiences. The system can unlearn things in order to overwrite or relearn things whenever it's necessary for its goals, though this will be significantly more difficult than normal learning.

B. Social Reasoning

An AGI would benefit from having innate ethical models based on Utilitarianism [16] and Deontology [17]. This system would retrieve the current context/state to evaluate and project its internal reward system onto other agents in its knowledge-space. In this manner, the AGI would be able to go beyond just theory of mind and empathize with other agents.

C. Planning

The planning component generates new sequences of states and scenarios from existing semantic and episodic knowledge. It uses the current state from short-term memory as a base to expand potential scenarios from. This component selects candidate predictions based on constraints/deadlines, estimated success (from previous attempts and similar scenarios), and randomly. Each prediction is evaluated using the motivation component (based on reward signals and appraisals based on goals). Plans can be represented in episodic memory as a series of predictions conditioned on the expected resulting states. Plans can be constructed/prioritized based on the likelihood of occurrence and the expected value if it does occur. These probabilities and expected values must be learned over time based on previous occurrences. Finally, plans are carried out based on either deadlines, real-time constraints, or reaching diminishing returns in the exploration stage of cognition.

VIII. Learning

Beyond just the acquisition and application of new knowledge and experiences, an AGI using high dimensional representations would likely benefit from other statistical learning methods.

A. Innate Learning Mechanisms

Learning over perceptual data, high dimensional representations, semantic representations, and episodic representations is automatic for these innate learning mechanisms. These mechanisms include the automatic acquisition of knowledge from experiences, the usage of acquired knowledge in retrieval/reasoning, causal learning, deep (Q) learning, decision trees, few-shot learning, transfer learning, and more. Since these learning mechanisms are not deliberate, they should be expected to converge quickly and with few data samples. This is where few-shot learning, transfer learning, and fine-tuning can be particularly effective for improving components which leverage DNNs. The enabling of few-shot learning can greatly reduce the number of samples required to learn novel concepts or skills.

There are a few more required functionalities in order to incorporate neural networks into our AGI design. Direct modification of DNN weights is a prerequisite to real-time model updates. Given progress in model pruning [], meta learning "warm-start weights" [], fine-tuning [], and transfer learning [], the prospects of directly updating models seems close. Another important functionality is the extraction of concepts and features from intermediate layers in DNNs. Concept bottleneck models [] and explainable AI approaches such as TCAV [] could provide the AGI with an ability to learn or enforce feature extraction with increased control.

B. Learning Strategies

Whenever learning becomes the primary goal of the AGI, (e.g., in studying, practicing, training, etc.) these learning strategies come into play. These include memorization/repetition, exploration of different learning architectures, altering retrieval and analogy, learning from other agents, imitation, planning, hypothesizing, and unlearning. The learning strategies employed can result in AGI systems which can allocate more computing power towards focusing and learning for the task at hand.

The AGI can also learn to create its own optimal representations and embeddings. Allowing the agent to alter its representations could improve the compression and approximation of information.

Other learning strategies include unlearning/pruning incorrect or unhelpful concepts in order to maintain an accurate and efficient knowledge base. While this may be more straightforward for static architectures, this poses a significant challenge for dynamic architectures. Removing information from the knowledge base and working memory is insufficient, as the DNN modules must unlearn/update concepts as well.

Finally, the AGI would benefit from learning when to regard knowledge from external agents as useful or useless. Heuristics indicating the importance of information from certain external supervision can be learned over time based on reward functions and previous success rates.

C. Leveraging Knowledge in Learning

In order to leverage knowledge in learning systems, the AGI could imagine and generate new scenario, similar to problem search but applied to learning. The recognition and retrieval of previous examples that are similar to the current instance or state being learned could help as well. Meta learning techniques would allow an AGI to apply concepts learned from a specific domain to learning itself. For example, if an agent learned about real-time scheduling concepts, it could apply these scheduling algorithms to its own studying behaviors.

D. Procedural Task Learning

All modules of our design can either be automatic/efficient or deliberate from learning, action, perception, attention, metacognition, emotion, planning, reasoning, retrieval, etc. Over time, these systems improve through practice, repetition, feedback. Here, the goal is to learn meaningful abstractions and approximations of knowledge. This system could be implemented as an indexes for faster querying and inference. This meta-level goal goes hand in hand with metacognition, where there is either a subconscious or conscious goal to optimize one of the design modules. These features and skills can still be transferable and abstractable to other things, as they will be represented using compositions of symbols. Eventually, anything can become learned as a system 1 automatic system, at the expense of reducing flexibility but improving runtime [18].

IX. Motivation

Our AGI's emotion and reward system is based on Appraisal Theory, where situational factors effect the produced emotion. They depend on the AGI's situations and goals. This can be simplified down to more complex version of a reinforcement learning model in which emotional appraisal is used as an intrinsic reward.

A. Goal Formation

While goals could be provided to an AGI as explicit priors, an AGI should be able to develop its own goals and subgoals. The functionality of goal formation would most likely be attributed to the AGI's personality and philosophical beliefs it subscribes to.

B. Emotion

By mapping appraisals onto the 8 basic emotions of joy, trust, fear, surprise, sadness, disgust, anger, and anticipation (and their varying levels of magnitude), the AGI can effectively communicate its emotions to other agents and empathize with other agents' emotions. Certain emotions could be innately associated with fast, instinctual responses such as fear and surprise. Whenever emotional appraisals appear to be endlessly spiraling into extremes, the AGI should take actions or reevaluate its current state in an attempt to revert to the baseline emotional state. This coping mechanism allows for the AGI to be grounded in the present state and remain receptive to incoming appraisals.

C. Reward Systems

Several innate reward mechanisms should be included into the positive reward function, including innate curiosity and correct inferences about the world. Negative reinforcement includes incorrect models/guesses, approximation uncertainty, disapproval from other agents, and ethical rule violations. Emotions can be simplified down to sentiment (positive-negative and magnitude) in order to be integrated into this reward system. These rewards ultimately bias the agent's likelihood to retrieve, consider, or implement certain actions, attention, goals, etc.

D. Decision Making

Decisions are passed from the planning component and are appraised and evaluated according to the current state in the short-term memory. Decisions may be impacted by certain extreme emotions, causing the AGI to react quickly or revert to the baseline emotional state. Decisions are represented as the mapping from observations to actions. Decisions have a variable amount of attractiveness based on the earlier defined reward mechanisms. An agent can eventually learn how metadata such as speed, efficiency, or cost, can impact decisions. Decisions can be associated with existing knowledge by being represented as episodic memory. The symbols composing a decision can aid in recalling previously made decisions in similar scenarios. Certain internal components can be processed as a decision, such as, attention, knowledge retrieval, procedural knowledge, reasoning, and meta learning.

X. Metacognition

There are several modes of metacognition in our AGI system [19, 20]. There is the continually invoked, behind-the-scenes metacognition involving resource and knowledge management. There is conscious metacognition, whenever the primary goal of the agent becomes metacognition and self-reflection. And there is heuristic-invoked metacognition which occurs whenever the internal secondary goals indicate that it is time to reevaluate certain system components. This AGI is able to run metacognitive processes on every other component, including metacognition.

A. Dynamic Architectures

In our AGI system, the architecture and each component must be somewhat mutable in order to improve performance via metacognition over time. In humans, there are often fixed structures and changing architectures. For example, knowledge acquisition, skill acquisition, and development over the course of a lifetime can result in drastic architecture changes. However, certain structures related to primitive actions, temporary storage, and performance may remain relatively fixed over time. Additionally, research in neuroplasticity suggests that cognitive capabilities in the form of structures and architectures are not fixed to a certain region of the brain [21]. These capabilities may redevelop over time if damaged. Ideally, according to our design principle, nothing should be hard-coded or inflexible. Every part of the system should be able to be self-edited or changed, with the exception of certain system 1 capabilities such as emotion/motivation, attention, reactions, retrieval, etc.

B. Continuous Metacognition

Continuous metacognition handles resource and knowledge management. This system leverages a preemptive priority-driven real-time scheduler for multiprocessing similar to those used in operating systems (global scheduling). The priorities are determined using dynamic scheduling based on the least amount of "slack time" available. This allows time-critical tasks such as responses, reactions, and knowledge retrieval to be prioritized. Longer tasks like conscious metacognition, imagination, learning, etc., will be postponed until computing power is available during a downtime. This type of metacognition uses metadata like computation time, deadlines, and goal rewards to determine the importance and priority of certain components and tasks. The processes of the metacognition component will typically take a long time, thus remaining a low-priority on this scheduler. Meta-level reasoning should have limitations on the depth of reasoning. This diminishing return would be encoded in the motivation mechanism to avoid the agent falling into the Homunculus Fallacy?

C. Conscious and Heuristic Metacognition

Conscious and heuristic metacognition will specify specific subcomponents and tasks to reevaluate. These types of metacognition use a digital twin of the current state/environment. This process can be likened to self-reflection and hypothesizing scenarios, decisions, and strategies. Heuristic metacognition occurs whenever certain heuristics raise red flags. This can be whenever high levels of uncertainty occur (high variance, few data samples) or when the motivation system is stuck in a feedback loop perpetuating negative rewards.

XI. Non-Human AGI Design

AGI can be useful for computer-related tasks that are tedious and humans do not have a good understanding for it (network security, privacy, web search, etc.) We provide a case-study for a non-humanlike AGI. This could be realized in a security system or online virtual assistant. There are several differences in the components necessary for producing general intelligence in these cyberspace environments/embodiments.

A. Levels of Generality

The most widely sought after goal for AGI is the creation of a human-level general intelligence. However, in the pursuit for human-level AGI, we may neglect simpler designs and architectures that enable generally intelligent behavior at lower levels. For example, generality of intelligence is highly dependent on the agent's goals. If the agent adaptively perceives, reasons, decides, and acts according to its goals, it is capable of simple general intelligence. For example, one could argue that mammals, insects, and even plants are generally intelligent agents. Intelligence is something that can also only be fully exhibited in a sufficiently complex environment. The environment must be rich and contain a variety of interactions to enable intelligent behavior to emerge (e.g., training a cognitive architecture on the MNIST dataset is insufficient in evaluating intelligence). However, with richer environments and data, an AGI must compensate with a greater amount of computational complexity, representational complexity, and sample efficiency. The question of general intelligence ultimately depends on the environment, embodiment, and agent goals. The final caveat to consider is that certain agents can behave intelligently at larger time scales, but may not appear intelligent at smaller time scales (e.g., hive/swarm, pencil+paper as external memory, individual neurons vs. human brains, etc.).

B. Cybersecurity

The environment a cybersecurity AGI experiences could be different than the real world. It doesn't necessarily need to be grounded in the same experiences, senses, etc. For example, the environment of a cybersecurity AGI would entail "sensing" network traffic, reducing the raw packet data into different types of concepts (DDOS packet, spoofed data, fuzzing attacks, etc.). The agent could attain knowledge from CVE reports, design documentation, datasets of attacks, etc. It could learn in real-time from red-team attacks, human supervision, and continuously learn in actual attack scenarios. Over time, the goal of this AGI agent would be to learn the appropriate responses and solutions to specific types of detected CVEs or network intrusions. It could intervene with actions such as blocking/modifying/sending network packets, notifying security engineers, or creating new CVE documentation. A secondary goal may be to either generate new attacks/defenses, or to attack a system as an artificial red-team agent. The AGI's reward system could be altered to be based on the trade-off between the safety/security, and latency of a network. In this embodiment social reasoning is unlikely to be needed.

C. *Privacy Assistant*

A virtual privacy assistant could greatly benefit from an AGI implementation. This assistant could read privacy policies, parse websites about companies and products, and attempt to infer exactly how data is used by companies. This privacy agent could learn from datasets, human supervision, language models, and more. Its goals are to help users with understanding by explaining nuances of privacy and presenting data in better ways. In this manner, it could decide to notify the user whenever their data is being collected via a certain activity or medium. This AGI would still need social capabilities and agent representations in order to determine user expectations and construct privacy contexts. Rather than representing concepts as just objects, agents, settings, and actions, it would benefit the AGI to add a "data" concept. Its reward system could be based on user feedback, intrinsic search/discovery/evaluation.

D. *Autonomous Vehicle (Fleet)*

Autonomous vehicles (AVs) are perhaps the most commonly discussed non-humanoid platform and use case which suits the capabilities of an AGI. An AGI would be embodied in an AV by reading sensor/camera/lidar/navigation data, making decisions about steering/acceleration/signalling, and communicating its decisions and processes to the passengers. As the environments an AV encounters vastly differs from the everyday environments and tasks of a human, it would benefit from constraining its knowledge space specifically to data related to driving, traffic, navigation, and vehicles. Many solutions addressing the individual problems within the field of AVs have been researched (e.g., object detection, lane-line detection, scene segmentation, object recognition). However, combining these vision capabilities with decision making, planning, reasoning, and actions (in real-time) is one of the most challenging problems facing AV researchers. Currently AVs are not capable of fully autonomous control over vehicles, and they must rely on human supervision/intervention. One of the less explored approaches to AGI in AVs is providing vehicles with basic autonomous capabilities and relying on vehicle to vehicle and vehicle to everything (V2V, V2X) communication. In this scenario, a generally intelligent "traffic controller" could send updates or decisions to each vehicle node. While the individual AVs would not be generally intelligent, intelligent behavior could emerge from the overall fleet of AVs.

XII. Other Cognitive Architectures

A. *Common Model of Cognition*

Researchers seeking to derive a standard model for cognition looked at systems and approaches developed in AI, AGI, cognitive science, neuroscience, and robotics. The authors converged on a single common understanding of the mind. The common model of cognition (CMC) involves a hybrid combination of symbolic and statistical processing [12]. In this framework/consensus, the mind is represented by independent modules that have distinct functionalities (e.g., perception, motor functions, working memory, long-term memory). The mind processes information in a cognitive cycle/loop. Through this cognitive cycle, learning emerges through the creation of new symbolic structures. While it is unclear whether the CMC will give rise to AGI, it notably lacks systems for motivation and regularization. Other researchers have mapped the CMC to locations and functionalities found in the human brain. Overall, the CMC suggests that architectures are fixed, and the CMC provides a baseline for critical modules that can be expanded upon. Our design can be seen as an extension on the CMC, where modules such as metacognition, motivation, meta-learning, and reasoning have been implemented. Metacognition and motivation serves to regulate the processes performed in each cognitive cycle and transform the CMC into a goal-directed model.

B. Symbolic Systems

Two of the most widely known cognitive architectures that rely on symbolic processing are Soar [4] and ACT-R [2]. Soar uses working memory as the centre for processing most of the data. Working memory is structured such that it handles situational awareness, goals, hypothetical states, and buffers for semantic/episodic memory in a graph structure. Soar also uses procedural memory to convert knowledge about skills, reasoning, and actions into rules. Soar can encode memories as either facts (semantic) or series of episodes (episodic). Many of the modules are parallelized, and chunking further improves cycle computation speeds while simplifying rules. Finally, perception represents visual data as 2D or 3D images from which features such as size, shape, color, and relations can be extracted. ACT-R is similar, although, the agents have task-specific modules to improve perception or storage. ACT-R suffers from several drawbacks such as a lack of metacognition and a lack of support for deleting working memory. Currently, a big drawback of these symbolic cognitive architectures is that they generally do not allow for easy transfer learning of skills/tasks [22]. Additionally, they cannot produce meaningful abstractions of skills and learned knowledge. As such, it is unclear whether the produced rules and decisions lead to a deeper understanding of knowledge. Finally, these cognitive architectures generally do not support the longer time-scales of cognition necessary for exhibiting generally intelligent behavior (¿ 10 sec band). We aim to support the acquisition of new cognitive capabilities through autonomous operation and self-awareness with our design. These are critical features for an AGI functioning beyond the scope of a few days.

C. Subsymbolic Systems

Several neuroscience inspired architectures have been created to accurately simulate brain functions and their low-level details. Spaun was created and evaluated with 2.5 million spiking neurons [23]. Using the MNIST handwriting dataset, they train the architecture to produce digits in similar handwriting styles, perform classification, and reason about the numerical positions and number lists. Leabra [24] directly bridges biology and cognition rather than using some abstract Bayesian model (DNNs) or highly detailed neural models (Blue Brain Project [25]). Their approach tries to combine the best of both extremes for cognitive modeling. Instead of using isolated models, these architectures rely on neural interconnectivity. However, some large-scale neural models face trade-offs to optimize for, such as balancing between stability and rapid updating, or balancing between pattern separation and sparsity of the model. Additionally, it can be quite difficult to fine-tune parameters and manually engineer constraints on large-scale neural models such as Leabra and Spaun.

D. Hybrid Systems

Hybrid cognitive architectures tend to combine approaches leveraged by both symbolic and subsymbolic architectures. Researchers have created SAL, a hybrid system which synthesizes ACT-R and Leabra. It combines these models to achieve multiple levels of abstraction [26]. Where ACT-R and symbol systems emphasize efficiency, tractability, and inspectability, Leabra and other subsymbolic architectures prioritize speed, capacity, and robustness. Combining these two approaches with different time ranges on Newell's level allows the AI agent to cover almost the entirety of all Newell's time scales cognition. This would allow for a robust system with increased parallelism, control, robustness, real-time capabilities, etc. Our approach matches most closely with SAL, although it substitutes Leabra with DNNs.

E. Deep Reinforcement Learning

Much research has been done to combine deep learning with reinforcement learning (RL) [27, 28]. These efforts have been successful in constrained environments where reward functions are easily define-able, although deep RL requires many more samples to achieve convergence. In deep RL, the agent interacts with the environment/world. It takes actions and receives observations. By combining additive rewards (e.g., payoff, gain, utility, costs), perception, reactive policies, and state representations, deep RL can produce an agent capable of reacting, planning, and implementing solutions in its environment.However, this approach is only sufficient for an AGI with a dynamic and flexible underlying architecture/structure. Additionally, this approach has yet to see success in highly complex environments such as real-world physics, and the problem of sample efficiency has not yet been solved. Defining constrained environments in the space of the real-world is challenging. Humans need to limit the scope of learning to specific tasks or environments (playing a game, learning from a course, learning a new job or task). The big differences are that 1) humans are more sample efficient and 2) humans can easily transfer knowledge from task to task.

In the field of AI and reinforcement learning, there is often a lot of emphasis on defining a good utility function to maximize. An example of this could be safety and reliability of arriving at a destination for autonomous vehicles. Humans are more complex, considering the future, the reciprocity from other agents, and acting on quick impulses and intuition. Ideally, an AGI design would account for this using more complex model building and self-regulating reward systems.

F. AI Generative Algorithms

The current approach in ML is to discover each of the pieces required for AGI with the hope that a group will combine all of those pieces. The approach of an AI-generating algorithm (AI-GA), proposes that AGI can be developed using generative and evolutionary algorithms [29]. Researchers define three pillars under which these algorithms optimize parameters: 1) trying to find the best deep learning architecture using meta learning, 2) trying to improve learning algorithms and task learning, and 3) trying to optimize and create effective learning environments. The most difficult challenge with this approach is constraining for the search space of each of the three pillars. The AI-GA approach would require an amount of computation that may not be available for an additional 10-30 years. Further, this approach requires a formal definition of general intelligence through evaluation suites and benchmarks which currently remain under-explored.

XIII. Evaluation

How can an AGI be evaluated for its capacity of intelligence? Some prior work have identified human intelligence tests which rely on a multitude of factors (e.g., perception, reasoning, learning, memory). However, AI systems have often been developed which can perform very well on specific intelligence tests without having any practical intelligence beyond these tests. We argue that measures of intelligence should be treated as benchmarks and “test set” data. Additionally, measures of intelligence alone are not enough to capture the extent of which a system is intelligent. Other important benchmarks include commonsense knowledge and reasoning and ability to complete general human-like tasks. Finally, proper benchmarking of human-like AGI should also involve a checklist of cognitive abilities present in humans.

A. Core Capabilities

These core capabilities represent the bare minimum structure required to support an autonomous AGI. These capabilities are quite similar to the common model of cognition or the common model of an intelligent decision maker [30].

- **[Perception]:** Sensing the external environment.
- **[Action]:** Performing actions in the external environment.
- **[Memory]:** Storing and accessing acquired knowledge.

- **[Motivation]:** Goals, rewards, and intrinsic motivation.
- **[Reasoning]:** Decision making, procedural knowledge, induction/abduction/deduction, planning, problem solving.
- **[Regularization]:** Regularization of the system using metacognition, “emotions”, scheduling, etc.

B. Cognitive Capability Checklist

These capabilities represent a more comprehensive list of which to evaluate the completeness of an AGI system.

1) Structures: These are capabilities and constraints that would likely be optimally implemented as explicit architecture components and modules.

- **[Symbols]:** Symbol representation and processing.
- **[Real-time]:** Operate in real-time.
- **[Continual Learning]:** Learn new knowledge and acquire new capabilities continuously and forever.
- **[Experience Replay]:** Timestamps on knowledge. Ability to “replay” episodic memories.
- **[Metacognition]:** Meta-learning, architecture changes, scheduling.
- **[Writeable Memory]:** Short and long term memory storage and management.
- **[Statistical Learning]:** Bayesian or Hebbian learning systems (ML and DNNs).
- **[Emotion]:** Appraisal-based or reaction-based emotions.
- **[Core Knowledge]:** Representations for basic building blocks of knowledge (physics, psychology, etc.).

2) Engineering Practicalities: These capabilities would be optimally implemented or acquired as knowledge priors attained through a training curriculum, pre-training, or engineering.

- **[Intention and Theory of Mind]:** Ability to represent and understand agents, their behaviors, and their intentions. Could be applied to others or self.
- **[Environment Richness]:** Ability to operate in a data and detail rich environment.
- **[Model Building]:** Constructing mental maps and mental models to represent concepts and scenes.
- **[Modal Representations]:** Use modality-specific representations and reasoning.
- **[Efficient Approximations]:** Ability to use universal value function approximators.
- **[Structural Organization]:** Ability to represent and leverage regularities, clusters, hierarchies, etc.

3) Behaviors: These capabilities should arise/emerge through the AGI’s continual learning in an environment and require little human supervision.

- **[Language]:** Humans must have ability to acquire syntax of a natural language.
- **[Numbers]:** Addition, subtraction, counting, recursion of unending natural numbers.
- **[Object Recognition and Mechanics]:** Understanding of physical objects. Reasoning about objects in space.
- **[Causality]:** Infants and adults understand physical connections between objects and motions.
- **[Abduction]:** Method of reasoning where you think hypothetically.
- **[Pedagogy and Social Knowledge]:** Teaching, learning, and social communicative knowledge.
- **[Perception]:** Perceiving and sensing environment.
- **[Mental Time Travel]:** Ability to plan for the future and reason about the past.
- **[Multi-Agent Interaction]:** Ability to cooperate, communicate, and coordinate with other agents.
- **[Flexibility]:** Behave flexibly as a function of the environment.
- **[Rationality]:** Exhibit rational or goal-oriented behavior.
- **[Learning]:** Learn from environment and experience. Acquire capabilities through development.
- **[Abstract Thought]:** Ability to reason about knowledge at any level of abstraction.
- **[Problem Solving]:** Ability to rationally solve novel challenges through transfer learning or trial and error.

- **[Planning]:** Predicting states, rewards, actions, scenarios, etc.
- **[Few-Shot Learning]:** Learning new concepts with extreme sample efficiency.
- **[Commonsense Knowledge]:** Understanding of most types of frequently used or foundational concepts used in human experiences.
- **[Diversity]:** Diversity in the types and levels of knowledge used.
- **[Self-Awareness]:** Ability to represent, reason, and alter state of self.
- **[Intelligent Exploration]:** Exploring, trial and error, and self play via intrinsic motivation.
- **[Unsupervised Learning]:** Operating and learning autonomously with little to no supervision.

XIV. Discussion

This section will discuss several counterpoints to the proposed design. It will also discuss the feasibility of creating, training, and deploying the proposed AGI system.

A. Limitations

A major problem emerges whenever AI agents are granted full autonomy – these AI agents engage in maximizing some utility function or metric without regard for much else. Systems must be regularized so that runaway processes or harmful goals do not occur/emerge. A good example of this has been seen in recommendation algorithms. Researchers have found that recommendation algorithms on social media maximize user time and attention, and as a result, these systems may recommend misinformation, extreme content, or content that induces outrage and other negative emotions. Those designing AGI must account for the ethical and safety issues that may arise from such autonomous control.

Another big shortcoming of our approach is the robustness and fragility of the system. We need a reliable method for quickly updating the DNN components and meta level self-monitoring loops to avoid common cognitive problems such as cognitive biases, over-learning, difficulty forgetting knowledge, reward addiction, excessive meta-cognition and rumination, etc. To address this, we can alter DNNs according to heuristics produced by the symbol system. Meta-level decision need to account for potential outcomes and uncertainty to determine the optimal decision.

B. Implementation

We plan to integrate DNNs as individual and dynamic submodules within our cognitive architecture. The overall architecture is centered around the knowledge representations and symbolic processing. For the symbol system, we plan to use a relational database (PostgreSQL) and a graph database (Neo4j, Embeddinghub, ArangoDB) to house long-term memory. Relational databases would excel for storing clusters, abstractions, groupings of concepts, and fast indexed searches. Whereas, a graph database would enable methods for ad-hoc search, discovery, analogy creation, and finding individual concepts. These entries, nodes, and edges will point to grounded representations (n-dimensional embeddings, raw data) computed from datasets or real-time data. Working memory will have to be implemented as a graph of embeddings and tensors (through which operations can be parallelized). We will use Python and Cython to prioritize ease of implementation over computation speed. We plan to map concepts to language tokens and integrate language models as modules for parsing and generating natural language. The cognitive architecture would likely need to be implemented on a device or cluster with at least many TB of storage, access to several GPUs, and a large amount of CPU cores/RAM. This server will receive and process a stream of inputs from the agent's embodiment. Initially, we will construct an AGI specifically to perform tasks related to online privacy and security in order to 1) limit the scope of the system implementation, 2) create an intermediate product capable of general intelligence in cyberspace, and 3) prove that DNNs and language models are capable of being integrated with a goal-directed cognitive architecture. This also removes a lot of the hard real-time constraints on the on the model. The next step of this implementation is to generalize its capabilities to human-like environments and embodiments.

C. AGI Ethics

One might bring up the concern of whether it is even ethical to develop AGI with reward functions that aren't traditionally rewarding to us (e.g., cleaning toilets, factory labor, etc.). An argument could be made that we do it to ourselves and to other people all the time. For example, at a meta-level, we might convince ourselves to find our work as more meaningful and enjoyable or redefine what we value and spend our time on. Schools also give rewards (good grades and job prospects) for performing well academically. We may even find that cleaning and cooking can become enjoyable or relaxing activities to us. Our view of ethics tends to evolve as we change and advance as a society. A bigger concern may be the economic implications of creating AGI capable of replacing many human jobs. Rather than applying AGI in a manner that replaces humans, perhaps we should develop technology that complements us. AGI can still play a very big role to play in this, especially when embodied in environments that are dangerous, unintuitive, or inhospitable to humans (e.g., cyberspace, extraterrestrial settings, factories, autonomous vehicles, disaster zones).

D. Future of AGI

With the rapidly accelerating growth of deep learning research and applications, the current prospects of AGI seem to be a secondary goal. For example, large transformer-based language models such as GPT-3 and PaLM have seen large success in being fine-tuned to solve specific tasks (e.g., DALL-E,). However, this approach alone will struggle to create goal-directed AI agents that require little human supervision and intervention once deployed. Deep reinforcement learning systems have been successful in tackling this challenge, albeit only in highly constrained environments such as tasks with structured rules and inputs (e.g., game-playing, programming,). The development of an AGI can further accelerate the adoption of AI in many tasks that require human-like intelligence. For example, the field of robotics and autonomous vehicle design has encountered significant challenges in creating safe, goal-directed AI agents that adapt to a wide variety of environments and scenarios. If we want well performing robots that can understand us and do a wide variety of tasks, AGI is the next big step in AI research. Beyond applying AI to new or complex tasks, AGI could significantly reduce the amount of human oversight currently required with training and deploying AI systems. Similar to large baseline models (Imagenet DNNs, language models, etc.), once trained, AGI would reduce the steep computational costs and professional manhours required to fine-tune models to specific tasks or environments. Additionally, a sufficiently robust AGI will likely generalize better than DNNs, and encounter fewer problems with fairness, safety, and security. However, the development of an AGI raises significant ethical issues: for example, AGI would likely be capable of automating many human jobs. This could require a significant restructuring of our economic policies.

AGI currently remains treated as an abstract faraway concept, and as such, there are incredibly few courses on AGI. Additionally, there are very few conferences and journals for AGI topics, and typical AI conferences not always suitable for AGI research. This may be a result of the incentive structure surrounding AI funding and research and the ambitiousness of the AGI vision. It doesn't make sense to tackle the problem of human-level AGI first. Rather than attempting to solve a borderline impossible problem, researchers must solve simpler problems and build up to human-level AGI. This could mean developing using a simpler environment, embodiment, or architecture.

As one of the original goals of AI, eventually AGI will have to be a major part of AI research. Narrow AI can only automate so many things. Additionally, if we want to exponentially accelerate production and technological growth, AGI is a prerequisite. While it may not be immediately obvious until an AGI is created, many significant strides towards AGI have been made. One can imagine the path to AGI as a progression with multiple axis. The AGI must exceed certain thresholds in the following categories – environment complexity, embodiment complexity, computational complexity, representation (abstraction) complexity, sample complexity. Below, we define five (+1) major milestones towards the creation of AGI:

- **[Milestone 0 (DONE 1951)]:** Birth of AI field.

- **[Milestone 1 (DONE 1993)]:** AGI that has the framework and basic capabilities to gather knowledge, make decisions, and act. Several cognitive architectures have been developed since 1993, and later, the rise of the common model of cognition.
- **[Milestone 2 (DONE 2015)]:** AGI with milestone 1 that can be successfully implemented in a complex environment and embodiment. It can successfully achieve its goals most of the time. Since 2015, several AI agents capable of knowledge acquisition and decision making have been developed in complex, diverse, yet constrained environments.
- **[Milestone 3 (ETA 2025-2030)]:** AGI with milestone 2, except it can be implemented in a wider variety of complex environments/embodiments. E.g., autonomous vehicles, general intelligence for multiple robot platforms, universal video game player, etc.
- **[Milestone 4 (ETA 2035-2050)]:** AGI with milestone 3 that can emulate a majority of human-level performance and behaviors. A "comprehensive" list of these capabilities can be found in Section XIII-B.
- **[Milestone 5 (ETA 2055-2100)]:** AGI with milestone 4 that can emulate all human-level performance and behaviors. It may also thoroughly surpass human-level performance in several of these capabilities.

XV. Conclusion

I hope this paper provides a good summary of the concepts discussed in the course and presents some fresh ideas that have not yet been thoroughly discussed. I thoroughly enjoyed taking this course and learning about AGI, especially since this is the one of the few (if not the only) AGI courses offered in the world. As someone who has always been fascinated with the concept of AGI, I am grateful I was able to learn so much. I'm excited to implement some version of an AGI in my research. Thank you so much for teaching this course, and I hope your retirement is relaxing and enjoyable!

References

[1] M. Richardson and P. Domingos, "Markov logic networks," *Machine learning*, vol. 62, no. 1, pp. 107–136, 2006.

[2] J. R. Anderson, M. Matessa, and C. Lebiere, "Act-r: A theory of higher level cognition and its relation to visual attention," *Human–Computer Interaction*, vol. 12, no. 4, pp. 439–462, 1997.

[3] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[4] J. E. Laird, "Extending the soar cognitive architecture," *Frontiers in Artificial Intelligence and Applications*, vol. 171, p. 224, 2008.

[5] R. Sun, "The importance of cognitive architectures: An analysis based on clarion," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 19, no. 2, pp. 159–193, 2007.

[6] S. McLeod, "Maslow's hierarchy of needs," *Simply psychology*, vol. 1, no. 1-18, 2007.

[7] S. S. Adams and S. Burbeck, "Beyond the octopus: From general intelligence toward a human-like mind," in *Theoretical foundations of artificial general intelligence*, Springer, 2012, pp. 49–65.

[8] P. R. K. Lindsay, *Understanding Understanding: Natural and Artificial Intelligence*. Scotts Valley, CA, USA: CreateSpace Independent Publishing Platform, Mar. 2012, ISBN: 978-1-46645058-5. [Online]. Available: https://www.amazon.com/Understanding-Natural-Artificial-Intelligence/dp/1466450584.

[9] K. D. Kinzler and E. S. Spelke, "Core systems in human cognition," *Progress in brain research*, vol. 164, pp. 257–264, 2007.

[10] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[11] A. Newell, *Unified theories of cognition*. Harvard University Press, 1994.

[12] J. E. Laird, C. Lebiere, and P. S. Rosenbloom, "A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics," *Ai Magazine*, vol. 38, no. 4, pp. 13–26, 2017.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] A. Kenny, "The homunculus fallacy," 1971.

[15] R. Seibel, "Discrimination reaction time for a 1,023-alternative task.," *Journal of experimental psychology*, vol. 66, no. 3, p. 215, 1963.

[16] J. S. Mill, "Utilitarianism," in *Seven masterpieces of philosophy*, Routledge, 2016, pp. 337–383.

[17] L. Alexander and M. Moore, "Deontological ethics," 2007.

[18] J. Laird and S. Mohan, "Learning fast and slow: Levels of learning in general autonomous intelligent agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[19] M. T. Cox, "Metacognition in computation: A selected research review," *Artificial intelligence*, vol. 169, no. 2, pp. 104–141, 2005.

[20] J. D. Kralik *et al.*, “Metacognition for a common model of cognition,” *Procedia computer science*, vol. 145, pp. 730–739, 2018.

[21] S. C. Cramer *et al.*, “Harnessing neuroplasticity for clinical applications,” *Brain*, vol. 134, no. 6, pp. 1591–1609, 2011.

[22] J. E. Laird, “An analysis and comparison of act-r and soar,” *arXiv preprint arXiv:2201.09305*, 2022.

[23] C. Eliasmith *et al.*, “A large-scale model of the functioning brain,” *science*, vol. 338, no. 6111, pp. 1202–1205, 2012.

[24] R. C. O’Reilly, T. E. Hazy, and S. A. Herd, *The leabra cognitive architecture: How to play 20 principles with nature*. Oxford University Press Oxford, UK, 2016, vol. 91.

[25] H. Markram, “The blue brain project,” *Nature Reviews Neuroscience*, vol. 7, no. 2, pp. 153–160, 2006.

[26] D. J. Jilk, C. Lebiere, R. C. O’Reilly, and J. R. Anderson, “Sal: An explicitly pluralistic cognitive architecture,” *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 20, no. 3, pp. 197–218, 2008.

[27] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[28] D. Silver, S. Singh, D. Precup, and R. S. Sutton, “Reward is enough,” *Artificial Intelligence*, vol. 299, p. 103 535, 2021.

[29] J. Clune, “Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence,” *arXiv preprint arXiv:1905.10985*, 2019.

[30] R. S. Sutton, “The quest for a common model of the intelligent decision maker,” *arXiv preprint arXiv:2202.13252*, 2022.