

2nd Semester AGI Journal

Brian Tang

2022-04-19



Contents

Journal Entries	4
Cognitive Architectures: Research Issues and Challenges	4
40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications . . .	7
A Standard Model of the Mind / Common Model of Cognition	7
Reward is Enough	7
The Quest for a Common Model of the Intelligent Decision Maker	8
Introduction to the Soar Cognitive Architecture	9
An Analysis and Comparison of ACT-R and Soar	9
Learning Rapid and Precise Skills	9
Neuroscience-Inspired Artificial Intelligence	10
A Large-Scale Model of the Functioning Brain	10
The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win!	11
SAL: An Explicitly Pluralistic Cognitive Architecture	11
Building machines that learn and think like people	11
AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence	12
Lecture Discussions	13
Lecture 14	13
Lecture 15	14
Lecture 16	15
Lecture 17	15
Lecture 18	15
Lecture 19	16
Lecture 21	17
Lecture 22	18
Lecture 23	19
Lecture 24	19
Lecture 25	21
Overall Trends and Issues with AGI Approaches	23
AGI Development in General	23
Symbolic Cognitive Architectures (Soar, ACT-R)	23
Brain Mimicking (Blue Brain, Spaun)	24
Deep Reinforcement Learning (AlphaGo)	24
Big Deep Neural Network (GPT-3, PaLM)	24

AI Generative Algorithms (Clune et al.) 24

Journal Entries

Cognitive Architectures: Research Issues and Challenges

- Should a cognitive architecture be static or dynamic?
 - Static architecture
 - * More efficient and constrained for optimization.
 - * Easier to code, monitor, understand.
 - * Lacks modularity and flexibility
 - Dynamic architecture
 - * Architectures as “learnable knowledge”. For example, to improve recognition/categorization at a meta-level, it create and use deep neural networks. Afterwards, the efficacy of this action could be represented in knowledge.
- Recognition and categorization:
 - Needs to be a system which can pattern match at every granularity. Take a visual perception system for example.
 - * An image of a scene can have objects arranged in a setting, and these might be matched with existing knowledge.
 - * Additionally, the objects within a scene each have their own sets of features (shape, color, motion, etc.). These must also be matched with existing knowledge.
 - * Some objects may have sub-objects which match with existing knowledge.
 - * (This system should be applied to other types of memories too such as decisions, plans, reasoning, actions, etc.)
 - How does one implement a computationally efficient method of pattern matching?
 - * We can compute “indexes” for recurring patterns to improve search speed. These indexes could map to existing episodic or semantic memories. They can be verified using classification or perceptual similarity.
 - * 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.
- Decision making:
 - How should decisions be represented?
 - * Observations -> actions
 - * Attractiveness of decisions based on reward/emotion
 - Agent can learn the utility of certain types of meta information like efficiency, speed, cost, etc. over time.

- How are decisions associated with existing knowledge?
 - * Can be represented within episodic memory
 - * Certain recalls based on symbols composing a decision.
- Perception and situation assessment:
 - How to construct and represent environmental situation?
 - * Have composition of symbols, each with representations grounded in different sensor modalities.
 - * Construct a single scene with setting, objects, agents, actions, etc.
 - * Save snapshots of these scenes, with emphasis on recalling different scenes.
- Prediction and monitoring:
 - How to choose the set of predictions worth considering?
 - * Select candidate predictions based on constraints/deadlines, estimated success (from previous attempts and similar scenarios), and just randomly.
 - * Evaluate candidate predictions using motivation component (reward signals and appraisal based emotion)
- Problem solving and planning:
 - How to construct and represent plans?
 - * Plans can be 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. The expected values must be connected to some reinforcement learning metrics.
 - When to carry out plans externally?
 - * Exploration vs. exploitation. Can be solved based on deadlines and real-time constraints.
- Reasoning and belief maintenance:
 - When is it useful to make inferences about existing knowledge?
 - * Whenever there is down time, 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.
 - What is the scope to consider when making inferences?
 - * The scope depends on the current/relevant goals of the agent.
 - How to construct the different types of inferences to be made?

- * Induction: Extrapolations based on past experiences.
- * Deduction:
- * Abduction: Construct/imagine hypothetical (functional) combinations of symbols/-concepts.
- How will the system regularize and balance the concepts it learns so as not to become overly extreme?
 - * Combining 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.
- Execution and action
 - How to represent and store skills/actions?
 - * As symbols themselves. Actions are concepts that are unique in that they describe relations between other concepts. They can be used as nodes as well.
 - * Actions and skills can comprise of subactions as well as the typical representational data forms.
- Interaction and communication
 - How to map stored knowledge to communicable natural language?
 - * Look into DALL-E and GPT-3. The way these models are created may provide some insight into expanding this to any modality (audio/noises, videos/actions, images/labels)
 - When to regard knowledge from external agents as useful/useless?
 - * Something that can be learned over time based on reward functions and previous success rates.
- Remembering, reflection, and learning
 - How to unlearn something?
 - * For static architectures, this is straightforward, but for dynamic architectures just removing it from the knowledge base or memory is not enough to undo the meta-level modifications to modules. These meta-level modifications should be recorded in metadata.
 - How to apply meta reasoning without falling into homunculus fallacy?
 - * Set limitations on depth of meta reasoning. Due to diminishing returns on reward and goal achievement.

40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications

- There are 40 years with almost 100 cognitive architectures. None of them produce a robust and intelligent (to the point of human-like intelligent) AGI. Why? What is missing? Is there even any one thing that is missing?
 - They are able to use perception, attention, action selection, memory, learning, reasoning, and metacognition, to perform tasks. Yet, they are still not up to par with human or animal capabilities.
 - The most practical and highly cited works are ACT-R, ART, and Soar.
 - One thing that seems to be unincorporated is multi-agent learning, where the cognitive architecture learns from direct interaction/supervision with humans. Many animals/humans require sufficient guidance/nurturing from adult supervision to form more complex thoughts and cognitive capabilities.
 - Perhaps there just isn't enough fast and parallel computing power to use cognitive architectures based off of the common model? Many of these architectures are based on closely emulating human cognition, but is this a flawed approach without the distributed infrastructure to back it up? Would a "hub and spoke" architecture design for processing be better? Is there any real benefit to using a dynamically self-changing cognitive architecture?

A Standard Model of the Mind / Common Model of Cognition

- Will the common model of cognition give rise to AGI?
 - AGI need some system for motivation (both intrinsic and external), but the common model lacks this functionality.
 - There is no apparent metacognition either.
 - The environment in which an agent is situated is just as important as the architecture itself. Environments not conducive to learning will make the acquisition and effective usage of knowledge more difficult.
- Procedural memory would have to incorporate a lot of different systems to create generalizable insights.

Reward is Enough

- Sure reward may be enough for an AGI with a dynamic and flexible underlying architecture/structure.

- This still requires a highly adaptive and self-referential knowledge system in the agent. This choice of the AGI designer’s representation of knowledge is challenging and should take reward into account.
- Also, sophisticated abilities may take a long time to arise from the maximization of simple goals, or a lot of human supervision.
- The agent still needs to create its own subgoals for solving a singular goal.
- Reward and emotion are highly intertwined. How do we convert emotion into reward? Is binarizing it into positive or negative sentiment sufficient?
- Why have reward give rise to these cognitive abilities when we already have systems that are able to do certain tasks well? Do we have to start from scratch in order to achieve “understanding”?
 - Is there some way to define some pre-trained intelligence abilities as primitives and request these based on reward?
 - * For example, some pre-trained image segmentation or object recognition model.
 - * Then, an agent can understand abilities and their relationship to the world at a higher level.
 - * Or maybe the agent can use model explainability and examine the different features that the model is using to perform the task in order to “understand” its abilities.
- Perhaps it would be useful to identify how to represent the different primitives of knowledge, as well as the primitive operations and capabilities that the agent can use?
 - Do these primitives have to be at the neuron or neural circuitry level? Or can we abstract this out to the cognitive band?

The Quest for a Common Model of the Intelligent Decision Maker

- The design seems to have a lot of overlap with the common model, with its uniqueness coming from designing the agent’s functions to evaluate and react based on reward.
- Why is there only an explicit reward signal embedded in the environment? Aren’t rewards and emotions in humans based on our beliefs, values, associations, and experiences? While humans are susceptible to issues like addiction and relying heavily on emotions and cognitive biases, an AGI created with only an understanding of external rewards will not be able to understand human emotions and motivations.
 - 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 than this, considering the future, the reciprocity from other agents, and acting on quick impulses and intuition. This decision making design does not seem to take these attributes into account.

Introduction to the Soar Cognitive Architecture

- The rules proposed by Soar do not seem to lead to a deeper understanding/generalization/abstraction.
- At what granularity are the rules and actions? What if the agent needs to backtrack or cancel its previously selected rule/action?
- There is not enough support for cognitive processing at higher levels of abstraction (e.g., metacognition on non-action capabilities, > 10sec band, complex tasks or events)
- There is a big emphasis on real-time capabilities and processing. There is a wide variety of human-level tasks (suitable for AGI) that do not require real-time constraints such as: virtual assistance, scheduling, and emailing, reading/understanding documents, etc.
- The design of Soar does not enable the acquisition of new cognitive capabilities, autonomous operation, and self-awareness. These are critical features for an AGI functioning beyond the scope of a few days. Additionally, it raises concerns about human supervision and scalability.

An Analysis and Comparison of ACT-R and Soar

- A big drawback of these symbolic cognitive architectures is that they generally do not allow for transfer learning of skills/tasks. Nor do they allow for abstractions. The chunking behavior acts more as procedural task learning, converting system 2 processes to system 1 processes rather than conceptual abstractions of tasks.
- ACT-R appears to have an advantage in computational cost by making all the modules parallelizable.
- In my opinion, forgetting (or at least overwriting) is a crucial part of learning and intelligence. In this aspect Soar has big advantages over ACT-R, especially when it comes to de-biasing or unlearning incorrect information. The robustness of an intelligence systems is dependent on this ability to forget.
- Representing knowledge in working memory only as a relational graph structure can have limitations, particularly for expensive searches.

Learning Rapid and Precise Skills

- Maybe because humans have a deeper understanding of the rewards and emotions involved with successfully completing the task, their retrieval systems are more motivated and can better leverage transfer learning.
- Seems like ACT-R requires a lot of manual coding and rule creation specific to the task. In principle, the framework can generally handle any task or environment, but this requires a lot of “designer effort”.

- I was not expecting half of this paper to be dedicated to describing the game mechanics...
- Skill acquisition is fast in humans due to lots of prior knowledge. Unfortunately, the prior knowledge experiment with ACT-R and the two games did not improve the sample efficiency. Meta learning approaches or learning pre-trained weights with DNNs do show promise for improving few-shot learning capabilities.

Neuroscience-Inspired Artificial Intelligence

- Not enough discussion on goal formulation/decomposition, metacognition, complex/causal reasoning.
- One of the big drawbacks of using deep learning that is not seen in approaches closer to actual neuroscience is the fragility of neural networks. DNNs are not very robust and are susceptible to attacks. If one were to construct an AGI using basic (undefended/non-pretrained) DNNs, one could cause the AGI to make poor decisions or cause it to learn detrimental/false information.
- It may be better to have explicit access to some of the intermediate features within neural networks (attention mechanisms, episodic memory, working memory) rather than letting the networks implicitly use these features.
 - Concept bottleneck models could be a promising direction (setting/extracting concepts/features from within a neural network).
- How do we extend DNNs to have continual learning capabilities? This suggests the need for some meaningful method of directly modifying weights/biases within an already trained network.
- Combining multiple DNNs into a single architecture/system is a nontrivial task. For example, how does the system interpret the embeddings/outputs in a way that is meaningful/transferrable to other DNNs? Maybe some analogies can be made about the structure of the embedding space or the uncertainty of the predictions.

A Large-Scale Model of the Functioning Brain

- How are the different brain regions initialized? Are these just blank slate neural circuits?
 - What about embodiment, training/learning phase?
- 2.5 million neurons is not nearly enough for human brains, although it may be sufficient for an insect or small mammal.
- It is a bit disappointing that MNIST was the input for Spaun rather than something more complex like Imagenet. It would be interesting to see whether Spaun could be used to create GAN-like capabilities or be used to perform Q&A or captioning tasks.

- Providing a side-by-side comparison of Spaun and DNNs would be very useful in determining which areas both architectures excel in (robustness, accuracy, learning rate, etc.).

The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win!

- There are a lot of mechanisms implemented/discovered by Leabra that deep learning can leverage (e.g., activation-based memory, hierarchical stages, updating vs. maintaining weights).
- Currently deep neural networks are very focused on associative learning, pattern recognition. There are limitations to cognitive
- Simulating all the functions of the brain without abstraction can get very computationally expensive. If you can approximate lower-level brain functions, you can save a lot of compute without compromising too much. E.g., you could iterate through all the individual steps of pruning, or you could drop out a random set of connections.

SAL: An Explicitly Pluralistic Cognitive Architecture

- This approach of combining neural architecture with a symbol system architecture is especially of interest to me. In my opinion, neural architectures would excel at covering the lower bands of Newell's levels, while a symbol system could be used to represent the higher level abstractions and cognitive capabilities.
 - The robustness of subsymbolic systems can be used to create more generalizable embeddings or representations that can be leveraged by the symbolic system.
 - Either you can wrap the subsymbolic systems within the symbol systems as representations, or you can run/train both in tandem for a given task/goal.

Building machines that learn and think like people

- Child machine vs. adult machine
 - “Child's mind as a notebook with rather little mechanisms and lots of blank sheets”
- If neural networks could encode objects, properties, forces, and dynamics, it could generalize to more complex physics.
 - Couldn't a symbolic reasoning system do this just as well without needing these concepts to emerge in the learning process?
- Motivates a system that can detect and classify things and concepts into core system groups: object, agent, setting, action

- There are a few more required functionalities in order to incorporate neural networks into an AGI:
 - Real-time model updates
 - Concept extraction and encoding in intermediate layers
- “Comparing the learning speeds of humans and neural networks on specific tasks is not meaningful, because humans have extensive prior experience”
 - This comparison is useful for developing AGI: learning speeds and prior experience are big factors to take into account
- We are reaching limits on sufficiently large training data and computing power... Simply increasing the size and training time is insufficient for developing goal-oriented AGI that can generalize to many tasks.
- There is model building at different granularities and levels of abstraction.
 - Model building could be as vague as constructing a scene of objects in a setting and their interactions.
 - It could also be fine-grained with building conceptual models of physics or mechanisms within an objects.

AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence

- “I believe it more likely that the AI-GA path will produce general AI first. That said, I have high uncertainty about that predication and think that either could get there first.”
 - I don’t think anyone has any good estimates about the timeline of developing AGI... Even with large amounts of compute, the AI-GA still needs to have primitives that eventually lead to all the building blocks of intelligence.
- “Assembling an AGI manually is a Herculean task requiring a large team”
 - Not all advancements in AI need to be implemented and combined individually. Some systems already encompass multiple of these components.
 - It seems for the manual approach, coming up with a general framework/architecture for adding components that interact with each other is the most challenging task. Filtering out the unnecessary or redundant capabilities comes next.
- The AI-GA approach has a few drawbacks. Particularly when combined with the environment generation approach:

- If we cannot understand an intelligence or behavior produced by an AI-GA, this could lead to challenges with integrating the AGI produced by the algorithm into society (e.g., sociopath AI).
 - * How are we supposed to evaluate an AGI produced by this AI-GA if there is no crossover in environments, worlds, etc.
- There are safety/ethical concerns with automating intelligence creation to this degree. Assuming the AI-GA approach works, how does one stop/prevent an AI-GA from surpassing human-level intelligence or ensure the AI-GA cannot be used for malicious purposes? Some sort of cognitive limitation should be placed to prevent this surpassing of human-level intelligence.
- “For example, the environment generator could be given access to the Internet.”
 - * This terrifies me... There are enough fairness issues with DNNs already. The Internet is not a great ground truth for our world (e.g., misinformation, hate speech, natural language biases, radical/polarizing content). If done in an unconstrained or unsupervised environment, this could create a problematic AGI. Examples include Microsoft’s Tay and harmful recommendation algorithms on social media.
- AI-GA approach can teach us many new things about encoding information, representing data, efficient learning schemes, etc.
 - Although, it will be challenging to come up with enough compute and constraints to produce an AGI through AI-GA...

Lecture Discussions

Lecture 14

- What are other possible architectural organizations for AGI
 - What are their tradeoffs?
 - How would a radically different AI embodiment affect this analysis?
 - * Must do multiple tasks in a dynamically changing environment.
 - No specification to link with an embodiment. Need some API?
 - Abstracts out perception and actions
 - Scene graph for environment, actions are defined by hardware, planners, etc.
 - Boston dynamics has defined tasks “go up the stairs”
 - If you can help it, don’t think about small level things
 - Shift from model-based to model-free

- We have the ability to index the different levels of hierarchical abstraction
- You have to learn the embodiment yourself
- Embodiment-specification functionality
- * What about other modalities?
 - Animals relying on scent. Tying in modalities in a new way. Octupi, macro-organisms (colonies/hives/swarms),
 - Swarm of drones.
 - How does this make the architecture different?
 - Capabilities are the same, but the connections between them might be different. Sub-agents need to be somewhat autonomous. Needs to be some macro-level goals.
- What capabilities/functionalities are missing?
 - A lot of architectures have nice siloed capabilities. Need less strict boundaries on what these concepts embody. Lots of redundancies in the brain and in the environment. Maybe having an idedic memory is useful.
- What capabilities/functionalities might not be necessary?
- What are other dimensions of evaluation that are missing?

Lecture 15

- How much does the CMC really say about AGI?
 - Is modeling human behavior the best way to approach AGI?
 - * CMC models human-like thinking
 - AI systems like DNNs are better at certain tasks
 - * If we have other technologies that do things very fast, how do we ensure that the internal models update?
 - If we learn something that challenges our models / partitions of knowledge, how do we update these mini models?
 - What are the components we're building out of?
 - What is the space of all potential common models?
 - Anything that doesn't have hub and spoke falls off empirically
 - * Searching the whole space for edges and subgraphs takes too long
- Most animals have these components found in CMC
- How do we handle perceptual-motor reflexes?
- Just looking at the data is not good enough, we need to establish some constraints

Lecture 16

- What are other points of comparison for CMC and CM-IDM?
 - Especially for AGI?
- Why do you think the CMC and CM-IDM analyses are a bit misaligned in terms of types of modules?
- Is reward enough?
- How do these compare to LeCun's model?
 - Abstract architecture, not a common model
 - Configurator decides based on tasks what architecture, learning rate, and other hyperparameters to use

Lecture 17

- It doesn't feel like we're on the verge of AGI, not in the next 10 years
 - What is missing or the constellation of things that are missing?
 - Once you have an agent and do task learning, how does it learn *on its own*?
 - * Soar will not be building abstractions of the world around it and further building abstractions related to actions and declarative structures in the world
 - There are very few systems that do this successfully
 - Not even those that can play Atari games and such do this
 - **Pure unsupervised abstract learning**
 - Big neural networks do this implicitly, but it never packages these implicit components specifically, label/cluster these concepts
 - * How do systems build this up and also interact with people?
 - Need to have a shared vocabulary
 - **What are the big holes in each cognitive architecture system?**
 - What goes into semantic memory?
- Humans start off with lots of neurons and start to lose the ones that aren't useful correlations.

Lecture 18

- ACT-R is not able to use natural language well, learn from the environment and experience, acquire capabilities through development, operate autonomously within a social community, be self-aware and have a sense of self, use core and commonsense knowledge, use emotion, use metacognition

- Deep RL does not have reasoning about the past, same with MDPs
- Is time just another property of objects?
 - * Color reasoning, space reasoning, etc. just requires current data
 - * Time reasoning requires some sort of a timestamp or other episodic representation
- Finer-grained ratings of AGI approaches
 - Architecturally universal
 - * Explicitly implemented, every agent does it
 - Architecturally supported and capable
 - * Aspects are available with architecture, system requires knowledge to use it
 - * Some systems have demonstrated plausibility
 - * Not clear that all aspects of this are possible with architecture as is
 - Architecturally agnostic
 - * Architecture is neither here nor there as to capability
 - * Has been demonstrated using knowledge in at least 1 system or many systems
 - * Unclear whether the aspects are possible with architecture as it is
 - Architecturally absent
 - * Architectural component necessary to support this is missing
 - * Knowledge used to implement it but would be better in architecture
- Putting things in architecture is useful for reducing computational complexity
 - Architecture is usually faster than having to search through N knowledge nodes

Lecture 19

- Which of the Newell constraints/capabilities are included in the paper?
- What additional constraints/capabilities would the authors of this article add to Newell's list?
- Judge NN's as presented in the paper on your combined list.
- Integration of nonreactive planning behavior.
- Merge sense of self and metacognition (metacognition is not just about self)
 - System 1 self-awareness, no control
 - System 2 self-awareness, deliberate
- Merge language and symbol representations
- Operate within a social community so it can engage verbally and non-verbally
- Make the list taxonomic and hierarchical

- Maybe we should do a bottom-up taxonomy creation of all the different constraints and features that human intelligence contains
- Abstraction/attention at different levels (isolating features)
 - Are these self-aware or reactive?

Lecture 21

- Create 2 lists: behavioral/functional and process/structure
- Sort into equivalence classes:
 - Necessary for any AGI
 - Necessary for mammal-level AGI
 - Necessary for human-level AGI
- Without this structure, will we be able to achieve reasonable runtime?
 - Maybe the things in functional, we don't recognize a structure for it yet...
 - Real-time is one of the strongest constraints
- Functional
 - Behave flexibly as a function of the environment
 - Operate in a rich, complex, detailed environment
 - Operate autonomously, but within a social community
 - Exhibit adaptive behavior
 - Operate in real time
 - Learn from the environment and experience
 - Use language, both natural and artificial
 - Be self-aware
 - Reason about the past and future
- Structural
 - Have a sense of self
 - Have representations for time
 - Use metacognition
 - Use emotion
 - Symbol representations and reasoning
 - Use modality-specific representations and reasoning
 - Use diverse types and levels of knowledge
 - Core and commonsense knowledge

- Any
 - Behave flexibly
 - Exhibit adaptive behavior
 - Use modality-specific representations and reasoning
 - Learn from the environment and experience
 - Operate autonomously
 - Operate in real time
 - Use diverse types and levels of knowledge
 - Operate in a rich, complex, detailed environment
 - Reason about the past and the future
- Mammalian
 - Core and commonsense knowledge
 - Use emotion
- Human-like
 - Use language
 - Use metacognition
 - Symbol representations and reasoning
 - Operate within a social community
 - Be self-aware and have a sense of self

Lecture 22

- Can natural language models leap over the need for symbol systems
- “If deep neural networks could adopt similarly compositional, hierarchical, and causal representations, we expect they could benefit more from learning-to-learn.”
 - How do we do this?
- “Deep learning models could incorporate these ingredients through some combination of additional structure and perhaps additional learning mechanisms, but for the most part have yet to do so.”
- No discussion about how to get real time performance
 - No discussion on the traditional model of task learning
- Other things to add to the list:
 - Different levels of abstractions

- Sample and computation efficiency
- Generalized task learning
- Building and processing causal models and representations
- Meta learning
- Compositionality

Lecture 23

- This AI-GA approach requires a formal definition of GI
- Amount of search required is ungodly/unfathomable
- Societies are critical to development of human-like intelligence
- Ignores the *hybrid approach*. Why not use neuroscience, AI, cog sci, chemistry, and physics to help constrain the searches???
- “Look where I’m pointing, don’t bite my finger” - Allen Newell

Lecture 24

- Define variations in AGIs
 - Different embodiments
 - * Hives/swarms
 - Each node is sensor + actuator
 - Similar-ish to neurons
 - Rules for interaction
 - Specialized vs. general agents
 - Smart vs. dumb
 - * Euclidean space with many different time scales
 - Our embodiment
 - Do opposable thumbs matter, long distance running, etc?
 - * Cyberspace
 - Lots of information is grounded in our world
 - Cybersecurity AGI (Mammal level)
 - Different levels of “general”
 - * Human-level general, do anything
 - * Granger reading (state machine, or stacks, or nested stacks of clusters of etc...)
 - Level of computations/abstractions available to agent

- * Generality is respect to goals
- * Mammal-level
- * Insect-level
- * Plants are generally intelligent (put through the 16 questions)
- * Intelligence is only fully exhibited if the environment is complex enough.
- * Superintelligence through runaway meta learning and lots of computational performance
- * We skip a lot of more simpler AGIs in the pursuit for human-level AGI.
- * ***Can create a grid with axis of general intelligence***
 - Environment complexity
 - Embodiment complexity
 - Computational complexity
 - Representation (abstraction) complexity
 - Sample complexity
- Maybe different Newell levels
 - * Certain things behave intelligently at larger time scales, which may not appear intelligent at smaller time scales (hive/swarm/pencil+paper as external memory). e.g., individual neurons vs. human brains
- Transform function/structure to a list of potentially necessary to support different types of AGI
 - How many of these functions actually necessary?
 - * Learning general intelligence vs general intelligence
 - * Functionalities make things faster (engineering practicalities)
 - * Core competencies/knowledge
 - * Modality specifics are practical improvements, just need some response to sensors
 - GI which is purely reactive and a lookup table doesn't need different types of knowledge
 - * Symbols are necessary
 - Why are some of these structures? Are they important? Maybe convert some of the structures into functions.
 - * Core/commonsense knowledge is for evaluation
 - When do you measure/evaluate/cull?
 - * Emotions are regularizers on behavior to prevent big mistakes and deadly mistakes from occurring.

Lecture 25

- DALL-E 2, GPT-3, PaLM, are all very cool things that aren't AGI, do we even need AGI?
 - YES - if we want well performing robots that can understand us and do a wide variety of tasks.
 - AGI can be useful for computer-related tasks that are tedious and humans do not have a good understanding for it (network security, teaching, privacy, etc.)
 - AGI is the next step after specialized deep learning AI. Solving problems like the autonomous vehicle challenge is really only *safely* possible through the creation of an AGI.
 - AI currently requires a lot of human oversight, professional manhours, and steep computational costs, which are all costly. An AGI that is robust will not be as fragile and require as much supervision as a DNN.
 - There is an argument against AGI though: the creation of AGI means many humans will be out of jobs and will require a restructuring of our economy.
- Looking forward, future of AGI
 - What are the milestones leading up to AGI?
 - * One can imagine the path to AGI as an n-dimensional chart. In this scenario, the AGI must meet these requirements.
 - Environment complexity
 - Embodiment complexity
 - Computational complexity
 - Representation (abstraction) complexity
 - Sample complexity
 - * Milestone 0 (reached)
 - AGI that has the framework and basic capabilities to gather knowledge, make decisions, and act.
 - * Milestone 1 (reached)
 - AGI with milestone 0 that can be successfully implemented in a complex environment and embodiment. It can successfully achieve its goals most of the time.
 - * Milestone 2
 - AGI with milestone 1, except it can be implemented in a wide variety of complex environments/embodiments.
 - * Milestone 3
 - AGI with milestone 2 that reaches or surpasses human-level performance and behaviors.

- * Milestone 4
 - AGI with milestone 3 that thoroughly surpasses human-level intelligence.
- What is the trajectory/evolution of AI over the future? Whether AGI is part of it or not?
 - * Eventually AGI will have to be a part of it. Narrow AI can only automate so many things. Additionally, if we want to exponentially accelerate production and technological growth, AGI is a prerequisite.
- Accessibility and teaching
 - There's no courses on AGI (this is the only one I could find)
 - Maybe that's because there's no incentive
 - Only 1 journal on AGI? Typical AI conferences not suitable for AGI?
- Timelining and AGI goals
 - Doesn't make sense to tackle the human-level AGI first, it's borderline impossible. Have to solve the simpler problems to build up to the human-level AGI. This could mean developing using a simpler environment, embodiment, or architecture.
- Defining constrained environments in curriculum learning (to discretize continual learning and infiniteness of environment)
 - Deep RL functions very well in overly constrained environments where rewards are well defined. 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.
- Is it ethical to program AI to have reward functions that aren't traditionally rewarding to us (e.g., cleaning toilets, factory labor, etc.)? Yes, we do it to ourselves and to other people all the time.
 - Examples: convincing/rewiring ourselves to find our work as meaningful and enjoyable, redefining what we value and spend our time on, schools giving rewards (good grades and job prospects) for performing well academically, even cleaning and cooking can become enjoyable/relaxing activities to us.
 - However, ethics tends to evolve as we change and advance as a society. Things that were socially acceptable hundreds of years ago are now illegal or widely regarded as highly unethical (eye for an eye, might makes right, slavery, eugenics, colonization, war, etc.).

Overall Trends and Issues with AGI Approaches

AGI Development in General

- Human-level AGI is much too ambitious of a goal. It can act as a good vision/pitch, but there need to be well-defined goals on the path to human-level AGI. E.g., start with a universal video game playing AGI, move to Unreal Engine physics simulations, move to constrained robot environments and tasks (autonomous vehicles, dog-level intelligence), finally adapt to human-level intelligence.
- Incentive structure for AGI is not there. AGI research does not fit well into the traditional “publish many papers” regime. It is also too ambitious to get easy funding for. Autonomous vehicles are a much less ambitious application of AGI that is able to get lots of funding (albeit still incredibly ambitious).
- There is very little discussion on AGI or cogsci in AI/ML courses. There are effectively 2 AGI course that I have been able to find (UMich, Berkeley). How can we expect to progress in AGI without teaching many students about it?
- AGI is realistically a multi-lab or company size effort (perhaps this is one of the reasons why DeepMind and OpenAI have garnered lots of press attention and success). Everyone needs to be on the same page with the AGI vision and dedicate 100% effort (Common model of cognition has helped with this).
- Not enough focus on AGI applications in cyberspace/internet (cybersecurity, virtual assistants, privacy).
- Not enough focus on composition, abstraction, causal reasoning, goal formulation, or metacognition in any of these systems. These are important for improving generalization to other tasks, environments, etc.
- Not enough focus on environments and data richness that are conducive to learning.
- Nothing has really shown continuous learning capabilities past the time-scale of months.
- Someone needs to look at all of these AGI approaches to figure out where/why they excel in certain areas, and combine the best capabilities of each system into one architecture.

Symbolic Cognitive Architectures (Soar, ACT-R)

- Does not leverage new hardware advances in parallel computing (GPUs).
- Too much focus on mimicking the real-time behavior of humans. Compute will probably not reach the level where real-time AGI is possible for another 5-10 years.

Brain Mimicking (Blue Brain, Spaun)

- Diversity of neurons and connections are still not fully understood.
- Even if the brain is simulated, no thought has been put into embodiment.
- Scaling to human brain size is unrealistic in their current implementations.

Deep Reinforcement Learning (AlphaGo)

- Sample complexity is too large.
- No transfer learning to other tasks.
- No way for system to define its own reward functions.
- Real-world is too big to have well-defined rewards. System needs to self-constrain the scope of learning to specific environments or tasks.

Big Deep Neural Network (GPT-3, PaLM)

- No actual implementation capable of creating goal-oriented behavior.
- Unclear whether lack of “grounding” in real-life experience will lead to misunderstandings.
- Requires millions of dollars to train the system.

AI Generative Algorithms (Clune et al.)

- Infeasible for at least the next 5-10 years, even if limited to just DNN search.
- No clear way to guide, evaluate, or understand AI-GAs.