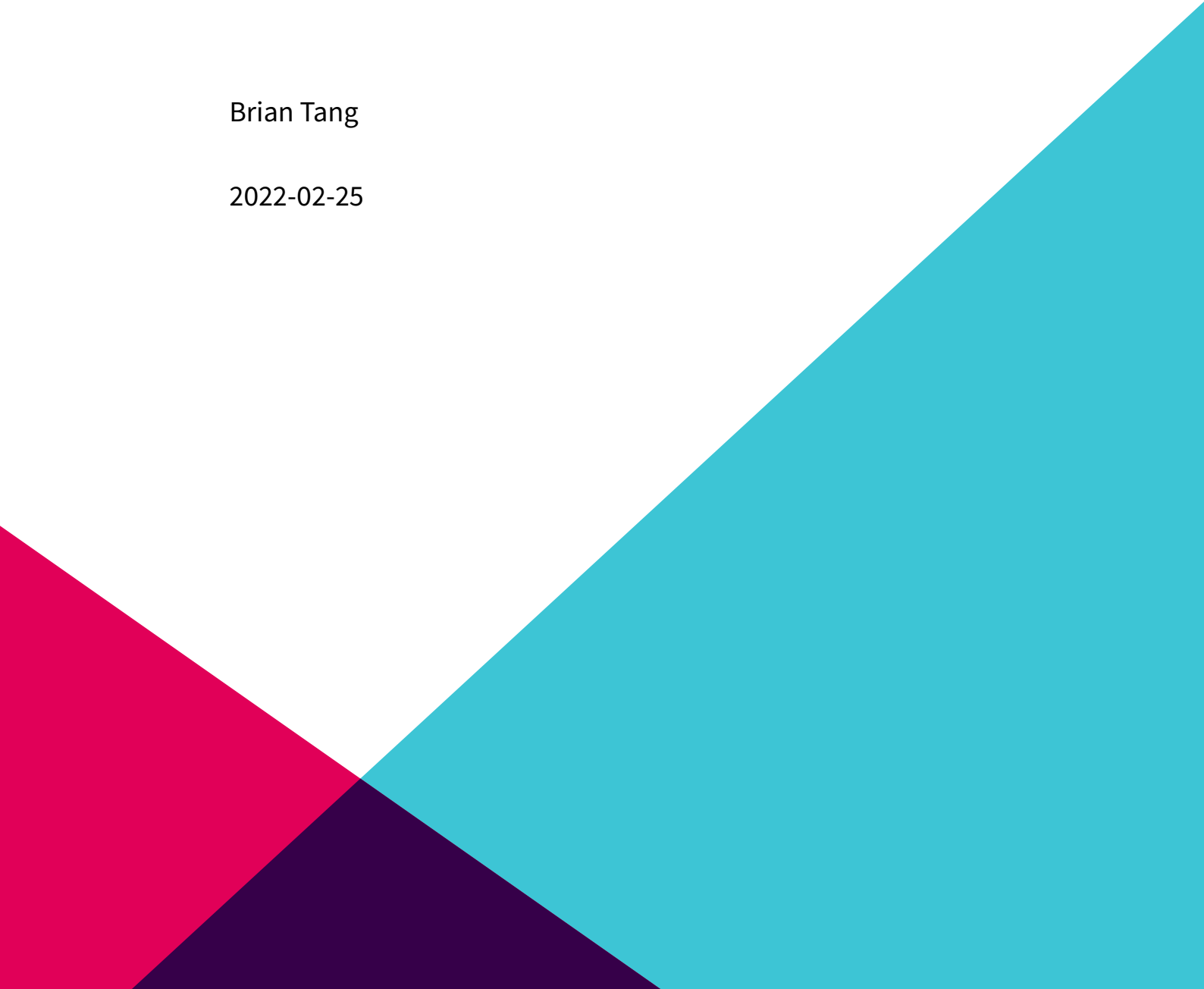

1st Semester AGI Journal

Brian Tang

2022-02-25



Contents

How to Navigate This Mess	5
Aggregated Journal entries	5
Lecture 1	5
Lecture 2	6
Lecture 3	7
Lecture 4	7
Lecture 5	8
Lecture 6	9
Lecture 7	9
Lecture 8	10
Lecture 9	11
Lecture 10	12
Lecture 11	12
Lecture 12	13
Lecture 13	14
2022-01-09 – AGI Readings 1	15
Universal Intelligence	15
Rationality and Intelligence	17
Lecture	18
Syllabus	18
AGI Introduction	20
Universal Intelligence	22
Journal	22
2022-01-16 – AGI Readings 2	23
Artificial General Intelligence: Concept, State of the Art, and Future Prospects	23
Intelligence, Knowledge & Human-like Intelligence	27
Journal	27
2022-01-16 – AGI Readings 3	29
Beyond the Octopus: From General Intelligence toward a Human-like Mind	29
Towards a bottom-up perspective on animal and human cognition	30
Artificial Intelligence and the Common Sense of Animals	30
Natural Intelligence	31
Lecture	32

Journal 35

2022-01-16 – AGI Readings 4 35

 Toward the quantification of cognition 35

 Lecture 37

 Journal 39

2022-01-19 – AGI Readings 5 40

 Foundations of Cognitive Science 40

 Human Cognitive Architecture 41

 Lecture 43

 Discussion questions 46

 Journal 47

2022-01-22 – AGI Readings 6 47

 Objects 47

 Agents 48

 Numbers 48

 Geometry 48

 Us vs. Them 48

 Other 49

 Lecture 49

 Journal 51

2022-01-26 – AGI Readings 7 52

 Embodiment of concepts 52

 Intelligence without reason 52

 Grounded cognition 54

 Lecture 55

 Discussion questions 57

 Journal 57

2022-01-28 – AGI Readings 8 58

 The computational origin of reasoning 58

 Commonsense Reasoning and Commonsense Knowledge in AI 59

 Lecture 61

 Discussion questions 64

 Journal 65

2022-02-02 – AGI Readings 9	65
Understanding	65
What does it mean for AI to understand?	66
Dark, beyond deep: A paradigm shift to cognitive AI with human-like common sense	67
Lecture	69
Discussion	71
Journal	72
2022-02-03 – AGI Readings 10	73
Computational models of analogy	73
Abstraction and analogy-making in AI	74
Lecture	76
Journal	79
2022-02-10 – AGI Readings 11	80
Metacognition for a common model of cognition	80
Metacognition in computation: A selected research review	83
Lecture	84
Discussion	88
Journal	89
2022-02-11 – AGI Readings 12	89
On the functional contributions of emotion mechanisms to (artificial) cognition and intelligence	89
Emotion in the common model of cognition	91
Emotion and Decision Making	92
Lecture	93
Journal	96
2022-02-16 – AGI Readings 13	97
Learning fast and slow: Levels of learning in GAIA	97
The computational gauntlet of human-like learning	98
Lecture	99
Discussion	102
Journal	102
Links	103

How to Navigate This Mess

- The aggregated journal entries are in the next chapter.
- Each lecture instance has a corresponding chapter. This entry is labeled with a date and lecture number.
- Each chapter contains comprehensive notes for each paper. These are infrequently sprinkled with my inserted questions, criticisms, and thoughts.
- Each chapter has lecture notes. These contain both discussion notes and lecture concepts.
- Finally, each chapter has a journal entry dedicated to ideation, questions, criticisms, and more “bigger-picture” thoughts.
- For each chapter, the discussion notes and journal entry sections contain the most important and “distilled” content.

Aggregated Journal entries

Lecture 1

- Criticisms of specialized intelligence
 - Identify challenging narrow tasks requiring lots of data and ERM in order to pass intelligence tests and benchmarks (Turing test, compression test, linguistic complexity).
- None of these definitions and such matter unless we come up with a form of practical generalized intelligence that is achievable given current computational constraints.
 - Why not just use a plethora of existing task sets, datasets, and real-world deployment to test for intelligence?
 - None of these intelligence tests factor in the concept of explainability nor the creation and understanding of abstract concepts.
- There are so many unexplored tradeoffs in the space of artificial general intelligence (no free lunch):
 - Compression of information and knowledge
 - Computational complexity of representative symbols
 - Real-time computing constraints
 - Single embodied vs. single non-embodied vs. cloud multiple-embodied agents
 - Computation spent towards metalevel reasoning
 - Utilitarianism vs other models of ethics

Lecture 2

- “AGI is impossible with realistic resource constraints” - I dispute this for the following reasons:
 - There is no formal proof that this is indeed the case, merely a conjecture
 - Narrow AI is becoming more and more computationally and memory efficient
 - Compression or distillation of knowledge is possible in both narrow and general AI
 - An AGI made of assembled components would meet intelligence tests
 - Recent advancements in deep learning, knowledge graphs, and common sense / causal reasoning prove promising for AGI
- Against hybrid learning: brittleness and weak robustness, brain uses unified infrastructure for a good reason
 - Isn't this more due to Occam's Razor of evolution?
 - Neurons are very adaptable and overall the system is robust, but we don't have enough resources to do this...
 - Immune system is an example of a very complex (component-wise) biological structure that is able to tackle a wide variety of “enemies” and adapt to handle future challenges.
- My envisioned AGI components (neuroplasticity)
 - Meta level learning and component management (IMPORTANT/CHALLENGING)
 - * Real-time system constraints
 - * Resource and component management
 - * Connecting the dots and extracting/abstracting
 - * Includes attention
 - Perception (Somewhat solved)
 - * Includes language, visual, audio
 - Action (Somewhat solved)
 - * All potential outputs (move, speak, more planning, more thinking, reflection, etc.)
 - * Includes communication
 - Memory (Unclear)
 - * Compresses, distills, and stores everything
 - Reasoning (Challenging)
 - * Includes causality, coincidence, deduction, induction
 - * Includes trial and error, navigation, planning
 - * Includes creativity, mental models, awareness of self and others (entities)
 - Learning (Somewhat solved)

- * Reinforcement learning
- * Pattern learning
- * Meta learning
- * Improving learning
- * Evolutionary learning
- * Online learning
- Motivation (Challenging)
 - * Goal formation, goal understanding, utility measurements
 - * Includes emotion

Lecture 3

- Humans and animals share similar underlying components and structures (neurons, brain sections)
 - However, there are sets of components that are unavailable to animals. There is likely some innate difference in either the structure or plasticity of our brains. Is this due to theory of mind, or some innate prior?
 - I.e., we can rewire and optimize certain parts of our brains for things we want to focus or specialize on.
- Is it possible that we're missing socialization in our AGI systems?
 - To go up the ladder of intelligence, we need basic understandings of trust and cooperation
 - After teaching 1 AGI, it can propagate the knowledge to others
 - We do not want to create AGI sociopaths which do not understand the meaning of empathy, trust, and cooperation
- Can we evaluate the progress of AGI using conceptual benchmarks?
 - E.g., does it understand object permanence, can it figure out what things it can interact with, does it transfer knowledge to other tasks, etc.
 - Might be a good idea to just create a gigantic checklist of components and go down the line from easiest and most important
- Use an algorithm to find the most memory and computational efficient architectures or technique for a given task

Lecture 4

- "Human minds are nested-stack automata, intrinsically computing indexed grammars. Our

natural languages clearly reflect this;”

- While this is likely true at the lower levels of computation, it is unclear how larger structures and areas of the brain interact.
- Why do humans lack exact memory?
 - We often lose connections to memories or add conflicting memory nodes.
 - Do we compress our memories?
 - * Could this have something to do with memory being episodic and experience based?
 - Is there some trade-off between learning capability and memory capacity?
 - Building something with unreliable memory forces us into different worlds.
- To achieve the effectively infinite memory needed for the discussed finite state machines, maybe compression is a key
- We should not assume that human-inspired cognitive or neuroscience components are feasible or the most practical for an AGI. While evolution and darwinism might suggest this, there has only been 1 species that has successfully achieved complex intelligence and human all-and-only capabilities. Most of evolution has focused on the ancestral, core, and developmental capabilities (innate, movement, perception, associations).
 - It may be unreasonable to assume that modeling neural circuitry is a scalable and practical approach for computers.
- Grammars and DNN structures
 - Transformers
 - * You have a sequence with associated features (vectors)
 - * For every point in the sequence
 - Query, key, value
 - Use query to pick another key which gets a value
 - Can do N queries (heads)
 - Allow for complex abstractions

Lecture 5

- Assuming that humans are symbol systems, we must have very vast memories and very accurate and efficient lookup systems
- “There is no way for a social group to assemble all the information relevant to a given goal, much less integrate it.”

- Not sure this is true given the context of human history, trust, and cooperation. Albeit slower, there is still a gathering of knowledge, consolidation of goals, and group reasoning to achieve tasks.

Lecture 6

- Core systems
 - Thing
 - * Each thing has a set of properties and interaction/relations with other things. These things can be represented by symbols.
 - * Objects
 - Physical interactions with other objects (containers, momentum)
 - Color, shape, usage, etc.
 - * Agents
 - Theory of mind
 - People, cats, dogs, (plants?)
 - Intentions, goals, actions, etc.
 - * Places
 - Setting in which objects and agents reside in.
 - Location, time, size, cleanliness, etc.
 - Core systems can be used to represent episodic and semantic memory easier, by having some structure for each thing/concept.
- Essentialism
 - Can boil down the essential properties of each core system to create a simplified and generalizable representation of a thing.

Lecture 7

- Rather than having innate priors of concepts, do we formulate them ourselves using what we perceive and find patterns in the world/environment around us? For example, why is it that the first few concepts children learn are shapes and colors? Is this due to our curriculum or is this the extent of patterns that they are able to grasp? How do more complex concepts build on top of existing ones? Is there some ordered hierarchy of abstractions/concepts?
 - Perhaps the concepts build upon one another, and we are able to imagine/generate new concepts using the foundational concepts and properties.

- Even for abstract concepts that seemingly have no modality, e.g., hope can be represented by a warm/positive feeling for the future.
- “If concepts were solely based on perceptual experience and passive associative learning, it remains unclear how children learn to distinguish between concepts (e.g., living vs. non-living) that are based on perceptually similar experiences (e.g., a living bird vs. a stuffed bird).”
 - I can see both sides of this argument. A child can perceive a living bird moving, flying, and chirping and over time passively associate these qualities with life. One could conceptually differentiate between the two using context clues from the environment as well (presence of trees, other birds, etc.) However, if a realistic stuffed bird were placed on a tree, one may use predictive modeling to discern finer-grained details via reinforcement learning. “My prediction was wrong. It was actually a stuffed bird. These are the minor differences to keep track of next time.”
- Subversion of expectations can provoke a strong emotional response (surprise) and inform future predictions.
- Is language just a convenient tool for fetching concepts and relations?

Lecture 8

- How do we automatically construct isomorphisms from the real-world?
 - We still need a way to map meaning to symbols via some sort of “environment modeling”. Can this be done using core systems and memory retrieval?
- How does learning using concepts work if they require prior knowledge of more foundational concepts?
 - There must be some system in place to begin picking up novel concepts via basic learning methods.
- Is commonsense reasoning more just simple/quick logical operations performed on existing knowledge in order to generalize rules to other situations/contexts?
 - Perhaps commonsense reasoning and knowledge can be seen as a way of warm-starting the learning process (not necessary, optional and helpful for learning/training speed)
- “of the year” vs “moldy blueberry soda”
 - Reminds me of GPT-3 and DALL-E (generating images from obscure captions)
 - * “Avocado chair”
 - * Possible reasons why this works very well:

- Multi-modal representations of concepts
- Very large network architectures (lots of “space” to fill up with knowledge)
- Very large amounts of data (lots of experiences to fill up the space with)
- Perhaps common sense knowledge is not something to be seen as a component or part of the AGI architecture, but something to be used as part of the learning process or curriculum for an AGI
 - Particularly useful for mapping language to concepts and for creating a system to categorize and add properties to objects/agents/places
 - * Then you can create more generalizations after having some modal data and context
 - Maybe this is why transformers work particularly well. Unfortunately, having multiple or 1 gigantic transformer is not realistic for run-time or memory.

Lecture 9

- Language is logic with statements?
 - I would say that it makes more sense that logic is done with symbols/representations of concepts
 - Language is more our ability to represent concepts in different modalities (talking/writing) and communicate it (to others)
- Maybe understanding requires consciousness - Does consciousness imply self-awareness? Or is it something “spiritual” that cannot be replicated.
- If understanding cannot be achieved unless the AI is embodied in our world, what happens if we “train” an AGI in our world (embodied in a robot) and transfer the AGI to a computer?
 - It could lose its capacity to reason/understand certain things without the ability to perform actions or perceive in its environment
 - How do we *safely* train an AGI or robot to interact with its surrounding?
 - * It could unintentionally break something or hurt someone, or worse, do it intentionally as a means of exploration
 - * Reinforcement learning and “scolding”?
- You don’t need technical understanding to use things, only functional understanding:
 - Form predictions of interaction results, feel bad if you get the prediction incorrectly, feel good if you get it correct
 - * Is this why gambling is fun/addicting?
- “Make instant coffee”

- Locate coffee jar, locate mug, open coffee jar, pour coffee into mug, check fluent state of mug, locate water, pour water.
- The order of some of these intentions/actions can be swapped around.
- Utilitarianism is not enough for forming choice preferences:
 - Need some hybrid form of ethics which at least combines deontological, virtue, and utilitarian ethics in some statistical framework.
 - * This system needs to learn new rules as well (deontology)
 - Otherwise we end up with AGI that sacrifice people to save more lives (surgery, trolley problem) or drawing incorrect conclusions (bounded rationality)

Lecture 10

- What is the difference between analogy and retrieval?
- Storage space is much less of a constraint compared to computing power when translating the capabilities of human brains into computers.
 - Brains can store around 2.5 petabytes of information
 - * A server which costs about \$100,000
 - GPT-3 has around the same number of neurons and synapses as the human brain (80-100 billion neurons)
 - * A cost of around \$10-20 million to train.
 - One approach could be storing highly detailed representations, but ignoring some of these details upon retrieval/analogy. This would be akin to abstracting away the finer details and would speed up computation.
- Analogies can be seen not just as mappings, but also as shortcuts for connecting concepts together.

Lecture 11

- Represent hypothetical task state and modify without interference with base-level representation or reasoning
 - For planning, imagination, etc.
 - Need to develop an imagination space separate from actual space
 - * Like a digital twin of the real world

- * Knowledge and learned interactions still need to be somewhat hooked up to actual systems (transfer learning/knowledge)
- What are other examples of metacognition?
 - Conscious scheduling of tasks
 - Management of goals
- How to apply to neural network models?
 - Meta-learning, hyperparameter tuning, automatic weight generation, neural architecture search
- Monitoring computations also takes time and cost
 - Motivates a real-time scheduler?
- There seems to be a similar overall structure in many of these meta-level systems (metacognition, emotion appraisal theory, predictive inference, learning systems). Generally it seems to be: perceive environment -> store knowledge -> evaluate internal state -> make decision -> perform action -> reevaluate environment and internal state.
 - Maybe we can leverage this by creating only 1 overarching system that does this perception action loop.
 - Then we can wrap the other meta-level systems around this

Lecture 12

- Can computers “feel emotion”?
 - Our brains send signals to distribute chemicals depending on our intuition and appraisal of a situation. These chemicals make the brain more receptive to pleasure (joy) or make the perception system in the brain more alert (fear).
 - Computers can likely emulate this experience through appraisal and reward systems. Computers may also have sensors that can interpret power usage, heat, fan speed, clock speed, etc. This could be seen as analogous to human’s heart rate, sweating, etc.
- Why have emotions in AGI?
 - Grounding in human-like emotional experiences.
 - Ability to efficiently interpret and convey emotions to/from humans.
 - Cultivate trust in AGI - human interactions (don’t want to create AGI sociopaths)
 - * Ability to empathize with humans and other AGI (crucial for social interactions and ethical reasoning)

- Emotions have to function in real-time
 - In order to properly influence reward mechanisms and reinforcement learning.
 - This means there should be a fast/intuitive appraisal mechanism and a slower reasoning appraisal mechanism.
- Beyond just appraisal theory
 - It helps people to further abstract/simplify emotions into either positive or negative (and magnitude).

Lecture 13

- “Expertise is applied piecemeal (1 element at a time)”
 - However, the lower level systems are incredibly parallelized. So, the level of focus/parallelization really depends on the cognition band level.
- “The dangers of sampling bias are well known”
 - This is a feature not just unique to neural networks or statistical learning, but also to humans!
- “Airplanes do not fly like birds, so why should computers think or learn like people?”
 - I agree with this. Why do we need to “start from scratch” when creating an AGI? Can we not leverage the advances in statistical learning to jumpstart the process? For example, one can use a pretrained DNN to handle perception-based tasks such as scene segmentation, object detection, classification, NLP, etc. It is unreasonable to restrict ourselves to creating an AGI that needs 5+ years to learn.
 - Computers are also fundamentally different from people, so we should leverage advantages and disadvantages that computers have (lots of memory and storage, fast operations, not great at parallelization)
- Is it unreasonable to try and incorporate deep learning and machine learning as low-level *components* of a human-level learning/AGI system?
 - These would be neural circuitry or cognitive level
- How to learn fast?
 - Use pretrained networks and few-shot learning.
 - Transfer learning from one domain to another.
 - Symbol systems, operations, and rules are fast.

- How to leverage knowledge in new learning?
 - Imagination and generation of new scenarios (like problem search but for learning).
 - Apply concepts learned from one domain to meta learning (e.g., see people who ask questions get praised -> ask more questions).
 - * Another example: learn about real-time systems -> use real-time scheduling strategies when studying -> learn faster
 - * These are more level 2 learning
 - Recognize and retrieve previous examples that have some similarity to new instance being learned

2022-01-09 – AGI Readings 1

Universal Intelligence

- Components
 - Verbal comprehension, word fluency, number functions, spatial visualization, associative memory, perceptual speed/reasoning
 - Analytical, creative, practical intelligence
 - Guilford's structure of intellect (120 categories)
 - Multiple intelligences (linguistic, musical, logical-math, spatial, bodily, intra/inter-personal)[Correlated by "g-factor"]
 - Fluid intelligence (general ability to deal with problems/complexity)
 - Crystallized intelligence (experiences)
 - Learning/adapting purposefully/effectively with environment to achieve a goal successfully
 - General mental capability involving ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, learn from experience.
- Agent-environment framework
 - Perceptions, actions, goals
 - Reward signals
 - Agent performs actions in an environment
 - * Obtains rewards and observes findings
 - Perception, action, reward spaces
 - * Uses symbols from some finite set
 - Short term rewards vs long term rewards

- * Needs to be able to plan far into the future
- * Long term reward parameter should be variable depending on the task or goal
 - Look into the future proportional to its current age
- General intelligence should handle a very wide range of environments
 - * Infinite environments, cannot simply do a uniform distribution
- Agent types
 - * Random agent (uniform random choices)
 - * Very specialised agent (good at a few specific tasks)
 - * General but simple agent (basic learning, takes action with highest expected reward in the next cycle)
 - * Simple agent with history (generates rewards from environments that are similar to previous ones)
 - * Simple forward looking agent (can look forward several steps and see opportunity costs)
 - * Very intelligent agent (perform well in simple environments and in more complex environments)
 - * Super intelligent agent (pick the best future reward action)
 - AIXI (AGI)
 - * A human (identifies simple structure of environments and how to exploit to maximise rewards)
- General intelligence tests properties
 - Need a definition of intelligence that is valid, meaningful, informative, wide ranging, general, unbiased, fundamental, formal, objective, universal, and practical
- Machine intelligence
 - A system that does well at a broad range of tasks
 - Ability to achieve goals in the world
 - Generates adaptive behavior to meet goals in a wide range of environments
 - Turing tests are not sufficient to establish intelligence, nor is it necessary
- Resource limitations to the definition of intelligence is wrong?
 - Definitely a factor to the *efficiency* of intelligence systems
- C-Test
 - g-factor view of intelligence
 - Ability to deal with complexity

- A lot of sequence prediction and abduction problems. A single unambiguous answer (1 hypothesis) for each pattern.
- Overcomes the problem of Kolmogorov complexity not being computable
- Assumes universal Turing machines can simulate each other in linear time
- Limited to passive environments, no dynamic environments
 - * Limited to simple symbolic pattern recognition and as such may not scale to complex real-world scenarios

Rationality and Intelligence

- Goal of AI is creation and understanding of intelligence.
 - Rational agency
 - Bounded optimality
- Need formal definition of intelligence
 - Perfect rationality
 - * Capacity to generate maximally successful behavior given available information
 - Calculative rationality
 - * Capacity to compute perfectly rational decision given available information
 - Metalevel rationality
 - * Capacity to select the optimal combination of computations and actions
 - Bounded optimality
 - * Capacity to generate maximally successful behavior given available information and computation resources
- Agents
 - Agent function specifies behavior under all circumstances
 - More closely related to game theory, control theory, evolutionary biology, etc.
- Rationality is relative to the agent's ultimate goals
 - Goal formation is not well understood
 - Someone moving to a city may consider many alternatives and tradeoffs and create a goal of buying a specific apartment
 - * They would focus their resources on achieving that goal
- Metalevel rationality

- Metalevel architecture to find an optimal tradeoff between computation costs and decision quality
- Calculate the information value of acquiring an additional piece of information through simulation or trial and error
- Monte Carlo tree search for game search
 - * Computation sequences
- Not all agent functions are feasible
 - Bounded optimality: do the best you can do with the resources you have
 - Act utilitarianism -> rule utilitarianism
 - Real-time computation constraints and deadlines
- Knowledge base can be improved by induction
- Reinforcement learning can learn information value and utility functions
- Compilation methods can improve faster decision making
 - Speed up problem solving
- Offline and online learning (lifelong learning)
- Metalevel Q functions

Lecture

Syllabus

- 1-3 papers per class
 - Expected to read all of them
- Class is about learning about what AGI is
- Keep a journal
 - Entry for every paper
 - 15 minutes of thinking and writing (not including reading)
 - About a paragraph
 - Send a pdf at mid-semester and end of semester
- 3 credit version
 - Analysis/taxonomies (discussed over the next 2 classes)
 - Final paper 10-20 pages
 - * What was missing in the class?

- * What conclusions can we draw?
 - * 1-2 person group
- Standard class
 - 20 minutes
 - * Paper presentation from professor
 - 20 minutes
 - * Group discussion #1
 - About 6 people
 - 2 key insights from paper
 - Usually answers to questions from professor
 - 20 minutes
 - * Group discussion #2
 - 20 minutes
 - * Full class discussion
- Class discussions
 - Read papers carefully
 - Write journal entry
 - Come up with questions and comments
- It is ok to have strong opinions
 - Need to have evidence, reasoning, or analysis to back up these ideas
 - * Always ask: why is this true?
 - * Be willing to change position
 - Evidence
 - * Empirical results, analysis
 - ***It is more important to be right in the long run rather than winning an argument today***
 - * We are looking for **truth**
 - Just because we have a system that does well for xyz, doesn't mean it will do well for abc...
- Course outline
 - Definitions of intelligence and AGI
 - Animals
 - Humans
 - Cognitive capabilities
 - Approaches for integrating capabilities in AGI

AGI Introduction

- AGI biggest obstacle
 - Desire for organization to be the winner and not to be right
 - * Momentum to ownership and being right
 - Does abstraction exist without meaning
- AI tasks
 - Game playing
 - Diagnostic imaging analysis
 - Autonomous driving
 - AGI is able to do all of these individual points
 - * Specialized AI is very good, AGI is not necessarily the best. Not everyone should do AGI
 - Could be that we never find the need for AGI
- Innate priors
 - “Loss function” of humans or AGI, innate goals
 - How does an agent know what better is without a notion of a reward
- Humans are not good at everything, but we have some ability to do anything
- AI activities
 - Most people working on applying/evaluating existing AI algorithms for new tasks
 - Improve and extend AI algorithms
 - Develop new AI algorithms for different tasks
 - Not many people working on AGI or human-level AI
 - PROGRESS IS NOT INEVITABLE
- Structure of AI (AGI is the whole thing)
 - Low level data processing
 - * Neural networks, symbols, bayesian networks, random code
 - Low level capabilities
 - * Matching, recognition, recall, tracking, learning, memory, decision making, attention, speech/vision processing
 - High level capabilities
 - * Meta reasoning, reasoning, planning, navigation, explanation, analogy, problem solving, language processing

- Applications
 - * Automated driving, automated assistant, stock trading, game playing, data analysis
- **How do we put these pieces together?**
 - * A cognitive architecture?
- AGI characteristics (similar to Goertzel's)
 - Has extended ongoing existence in an external environment
 - * Autonomous, no moment-to-moment supervision
 - * Constant automated learning by itself, and it has access to some external environment
 - Can pursue many different types of tasks
 - * Including novel tasks
 - Always learning and can generalize what it has learned
 - Incorporates many cognitive abilities
 - Interacts with humans and other agents
 - * Might not be through natural language
 - Not necessarily human-level or human-like
 - * We may want to build systems that do different things
 - AGI very good at network security
 - * Don't always refer to humans and animals for inspiration
- GPT-3 is amazing, but probably not AGI
 - Drunk old uncle whose had too much to drink, sometimes comes out with amazing true facts, other times it's absolute garbage
- Big questions
 - What are the defining characteristics of AGIs?
 - What are the structures, regularities in tasks and environments that can be exploited for general intelligence?
 - What computational capabilities are required/helpful to support/exploit those regularities in a general way?
 - Which of those capabilities are innate, which are learned?
 - How are the capabilities organized within an agent so as to be available when needed?
 - * What are the underlying computational architecture?
 - * The brain and AGI are not homogeneous

Universal Intelligence

- Shane Legg is founder of Deep Mind (2010)
- Marcus Hutter is a DeepMind senior scientist
- AGI conference is not a typical academic conference
 - Hobbyists, academics, etc.
- Human intelligence tests
 - Assume basic human capabilities (relative to normal, average humans)
 - Not good at distinguishing levels or breadth of AGI
 - AI systems have been developed that perform very well on specific intelligence tests (pattern matching, etc.)
- What is intelligence?
 - Conglomeration of specific capabilities
 - “Intelligence measures an agent’s ability to achieve goals in a wide range of environments”
 - Agents
 - * Agent has ongoing existence with multiple tasks over time, not onetime execution
 - * Interaction with environment via perception and actions
 - Environments
 - * What are important properties?
 - * Dynamic? How dangerous?
 - Goals
 - * Objective to achieve in environment
 - * Focus on reward
 - Reward is only about the current situation or a future reception of the reward
 - Assume ability to learn
 - * Reward alone misses ability to receive description of novel objective
 - Where there is a description of a hypothetical/future state can guide current behavior
 - * Doesn’t discuss intrinsic reward (extrinsic vs intrinsic rewards)
 - It obviously shouldn’t give itself dopamine every time

Journal

- Criticisms of specialized intelligence

- Identify challenging narrow tasks requiring lots of data and ERM in order to pass intelligence tests and benchmarks (Turing test, compression test, linguistic complexity).
- None of these definitions and such matter unless we come up with a form of practical generalized intelligence that is achievable given current computational constraints.
 - Why not just use a plethora of existing task sets, datasets, and real-world deployment to test for intelligence?
 - None of these intelligence tests factor in the concept of explainability nor the creation and understanding of abstract concepts.
- There are so many unexplored tradeoffs in the space of artificial general intelligence (no free lunch):
 - Compression of information and knowledge
 - Computational complexity of representative symbols
 - Real-time computing constraints
 - Single embodied vs. single non-embodied vs. cloud multiple-embodied agents
 - Computation spent towards metalevel reasoning
 - Utilitarianism vs other models of ethics

2022-01-16 – AGI Readings 2

Artificial General Intelligence: Concept, State of the Art, and Future Prospects

- General intelligence is a fundamentally distinct property from task or problem capability
 - AGI need not be infinitely generalizable
 - AGI can simply bridge the gap between current AI programs (narrow)
 - Often seen in robots, chatbots, etc.
 - Ability to achieve a variety of goals, tasks, in different context/environments
 - Can handle problems and situations different from those anticipated by creators
 - Good at generalizing the knowledge it's gained so as to transfer learn
 - Impossible with realistic resource constraints
 - * ^ *I dispute this*
 - Real-world systems can have varying levels of generality, inevitably more efficient at learning some things than others
- AGI Hypothesis

- Creation and study of AI with broad scope and strong generalization capability, different from narrow scope and weak generalization
- Not necessarily create human-like general intelligence
- Competencies
 - Perception (vision, hearing, touch)
 - Actuation (physical manipulation, tool use, navigation)
 - Memory (implicit, working, episodic, semantic, procedural)
 - * No introspection, awareness, 1st person experiences, facts/beliefs, sequence/parallel of physical or mental actions
 - Learning (imitation, reinforcement, trial and error, reading, listening)
 - Reasoning (deduction, induction, abduction, causal reasoning, physical reasoning, associational reasoning)
 - Planning (tactical, strategic, physical social)
 - Attention (visual attention, social attention, behavioral attention)
 - Motivation (subgoal creation, affect-based motivation, control of emotions)
 - Emotion (expressing, perceiving, interpreting emotion)
 - Modeling self and others (self-awareness, theory of mind, self-control, other-awareness, empathy)
 - Social interaction (social norms, communication about relationships, inference about relationships, group interactions)
 - Communication (gestural, verbal, pictorial, language acquisition, cross-modal communication)
 - Quantitative skills (counting, math, quantitative properties, measurement)
 - Building/creation (physical, conceptual, verbal invention, social construction)
- Cognitive architecture (SOAR)
 - Fixed structure for all tasks
 - Symbol system
 - Modality-specific knowledge
 - Large bodies of diverse knowledge
 - Different levels of generality of knowledge
 - Diverse depth of knowledge
 - Beliefs independent of current perception
 - Hierarchical control knowledge
 - Meta-cognitive knowledge
 - Spectrum of bounded and unbounded deliberation
 - Diverse, comprehensive learning

- Incremental, online learning
- Embodiment approach
 - Body-environment interaction makes intelligence easier understood
- AGI approaches
 - Symbolic
 - * Centralized control of perception, cognition, and action
 - * Markov logic networks, inductive logic programming, genetic programming
 - * ACT-R (production rules and connectionist dynamics)
 - * Cyc (natural language engine)
 - * EPIC (production rules and cognitive processor, symbolically coded features rather than raw sensor data)
 - * ICARUS
 - * SNePS
 - * SOAR
 - * Symbolic thought lets us generalize most broadly
 - * Symbolic AI architectures are incapable of giving rise to emergent structures and dynamics
 - Emergentist
 - * Abstract symbolic processing
 - * Good with patterns, reinforcement learning, associative memory
 - * DeSTIN (hierarchical pattern recognition)
 - * HTM (pattern recognition)
 - * SAL (based on IBCA, similar to hippocampus)
 - * NOMAD (neural darwinism, natural selection of neurons into configurations that work best)
 - * Ben Kuipers' bootstrap learning (reasoning and reinforcement learning)
 - * Deep learning
 - * Create a large set of simple elements capable of adaptive self-organization, since human brains are large structures
 - * How does the brain organize things into these groups of neurons *automatically*? The right architecture is needed, not the underlying mechanics
 - Computational neuroscience
 - * Lets just understand the brain better
 - * IBM Blue Brain Project
 - * Large scale brain modeling

- * Neurogrid project
- * Computers and brains are not directly transferable and compatible (computation restrictions)
- Artificial life
 - * Ecosystem is too computationally complex
- Developmental robotics
 - * Allow robots to learn via intrinsic motivation
 - * Juergen Schmidhuber
 - * FLOWERS, IM-CLEVER, SAIL
 - * Young children do this
 - * Robots are too crude and don't have good enough infrastructure
- Hybrid
 - * Need some combination of the above
 - * CLARION (NN, RL, etc.)
 - * CogPrime
 - * DUAL
 - * LIDA
 - * MicroPsi
 - * PolyScheme
 - * Shruti
 - * 4D/RCS
 - * Brain is a lot of different parts working together
 - * Gluing together a bunch of inadequate systems isn't enough to make an adequate system
 - Brittleness and weak robustness
 - Brain uses a unified infrastructure for a good reason
- Universalist
 - * AIXI
 - * Create some meta-algorithm to automatically figure out how to make AGI smarter
 - * Computationally infeasible, but ideal
 - * More rigorous than ad-hoc approaches
- Measuring AGI
 - Turing test
 - AIQ
 - Text compression (new compression methods)
 - University test

- Artificial scientist test
- Coffee test
- Measuring incremental AGI progress is hard, in part because it's all or nothing (6 components is not enough, but 10 components does the trick)
- AGI is still useful even though narrow AI exists (the next step)
- Showing a practically demonstratable AGI would get everyone super hyped and prove its usefulness

Intelligence, Knowledge & Human-like Intelligence

- Intelligence is pretty much rationality
 - Differs from adaptive intelligence, general intelligence, or human-level intelligence
- Innate knowledge
 - Intelligence is how well non-learning agents use knowledge to perform a task
- Task experience
 - An agent's knowledge increases with each interaction with the environment
 - Can retrospectively analyze and reflect when there's no time pressure
- Exploration
 - Seek out knowledge in its environment
 - Personal exploration in a world that is large and rich is so useful
- This definition allows meaningful comparisons between humans at similar ages as well as humans and non-humans
- Can highly intelligent human-like agents have different cognitive architectures?

Journal

- "AGI is impossible with realistic resource constraints" - I dispute this for the following reasons:
 - There is no formal proof that this is indeed the case, merely a conjecture
 - Narrow AI is becoming more and more computationally and memory efficient
 - Compression or distillation of knowledge is possible in both narrow and general AI
 - An AGI made of assembled components would meet intelligence tests
 - Recent advancements in deep learning, knowledge graphs, and common sense / causal reasoning prove promising for AGI

- Against hybrid learning: brittleness and weak robustness, brain uses unified infrastructure for a good reason
 - Isn't this more due to Occam's Razor of evolution?
 - Neurons are very adaptable and overall the system is robust, but we don't have enough resources to do this...
 - Immune system is an example of a very complex (component-wise) biological structure that is able to tackle a wide variety of "enemies" and adapt to handle future challenges.
- My envisioned AGI components (neuroplasticity)
 - Meta level learning and component management (IMPORTANT/CHALLENGING)
 - * Real-time system constraints
 - * Resource and component management
 - * Connecting the dots and extracting/abstracting
 - * Includes attention
 - Perception (Somewhat solved)
 - * Includes language, visual, audio
 - Action (Somewhat solved)
 - * All potential outputs (move, speak, more planning, more thinking, reflection, etc.)
 - * Includes communication
 - Memory (Unclear)
 - * Compresses, distills, and stores everything
 - Reasoning (Challenging)
 - * Includes causality, coincidence, deduction, induction
 - * Includes trial and error, navigation, planning
 - * Includes creativity, mental models, awareness of self and others (entities)
 - Learning (Somewhat solved)
 - * Reinforcement learning
 - * Pattern learning
 - * Meta learning
 - * Improving learning
 - * Evolutionary learning
 - * Online learning
 - Motivation (Challenging)
 - * Goal formation, goal understanding, utility measurements
 - * Includes emotion

2022-01-16 – AGI Readings 3

Beyond the Octopus: From General Intelligence toward a Human-like Mind

- Octopus learns very quickly and solves problems creatively
 - Short lifespan
 - Hatchlings must learn quickly and be lucky to survive
 - 300 million neurons
- Examples
 - Opening a screw jar
 - Using coconuts as shelters
 - Shooting out the lights with water
 - Spatial learning
 - Observational learning
 - Camouflage and behavioral mimicry
- Ladder of intelligence
 - Asocial reasoning
 - * Hide, forage, hunt, kill, flee, eat, what, where
 - Social reasoning
 - * Nurture, protect, feed, bond, give, share
 - Animal cultural reasoning
 - * Follow, cooperate, play, lead, warn, trick, steal, teach
 - Oral linguistic reasoning
 - * Promise, apologize, oath, agree, covenant, name
 - Literate reasoning
 - * Library, contract, fiction, technology, essay, spelling, acronym, document, book
 - Civilization-scale reasoning
 - * Democracy, empire, philosophy, science, mathematics, culture, economy, nation, literature
- Linguistic grounding
 - Meanings of words
 - * Based on basic facts of human bodies
 - * Used metaphorically

Towards a bottom-up perspective on animal and human cognition

- Basic building blocks of cognition is shared across many species
 - Do animals have theory of mind, culture, imitation, etc.
- Evolution acts on predispositions and motivations
 - Also retains core learning mechanisms
 - Face recognition in humans and primates are similar
- Mirror neurons are for instinctive actions and such
 - Is imitation understood at a meta level?
 - All imitation has a shared neural perception-action foundation
- Prosocial behavior and empathy
 - Where do altruistic tendencies stem from?
 - * Independent of incentives and long-term benefits
 - * Spontaneous helping
 - * Likely stems from empathy
 - Emotional contagion

Artificial Intelligence and the Common Sense of Animals

- Common sense existed before language
 - “The falling rock smashed the bottle”
 - * Understand motion and space for “falling”
 - * Know what a rock and an object is
 - * Understanding objects, motion, space, causality
- Deep RL
 - Embodied systems can interact and explore
 - Trial and error to maximise expected reward over time
 - DQN: deep RL
 - * Not data efficient, must play many games
 - * Brittle, not robust
 - * Inflexible, cannot be transferred to another game
 - Intrinsic motivation (curiosity)

- Training protocols
 - Sequential multitask training (episodic)
 - Life of animals is continuous
 - Structure tasks as a curriculum (simple -> complex)
- RL
 - Must perceive objects by exploiting correlations at the pixel level
 - * Doesn't understand that 3D objects occupy space
 - Containers and enclosures are also challenging
 - * Object permanence

Natural Intelligence

- Humans possess language, animals don't
 - They also possess theory of mind (meta-level understanding of our own minds as decision makers)
 - * Can project this model onto others
- Intelligence operates in **real time**
 - Governed by law of physics, safety, usability
 - Are these decisions always conscious? May or may not be
- Mammals have goals, animals are curious
 - Chimpanzees need to be first taught by humans (symbolic communication system), before understanding categories of objects
 - Tool usage (cause and effect with tools and goals)
 - Mentally rotate objects, plan object manipulation
- Clever Hans
 - Taught to count (but actually just memorized cues from handler)
 - Empirically studying animal intelligence is difficult
 - Emotions are an option for understanding
 - * Same with attention and gaze
- Core human cognitive abilities
 - Cognitive modularity (are abilities separable?)

- Language
 - * Humans must have ability to acquire syntax of a natural language
 - * Linguistic utterances alone isn't enough for children to figure out the language's syntax
 - Must be predisposed to consider certain patterns of grammatical rules and make use of non-linguistic information
- Number
 - * Addition, subtraction, counting (animals)
 - * Recursion of unending natural numbers (children)
- Object recognition and mechanics
 - * Children understand physical objects
 - They can also represent situations they do not actually see (imagination)
 - * Animals can deal with objects in space
 - * Infants and children suck at reasoning about objects' motion
- Causality
 - * Infants understand physical connections between objects and motions
 - * Animals do not possess causality
- Intention and theory of mind
 - * Agents' behaviors (particularly humans) are intentionally done to achieve goals
- Abduction: thinking hypothetically
 - * Method of reasoning where you think hypothetically
 - * Hypothesize an explanation
 - * Often leads to wrong explanations and a failure to consider other explanations
- Pedagogy and social knowledge
 - * Animals do not have a cumulative culture because they don't have cognitive capacities for representing, communicating, and recording the products of culture

Lecture

- Learning AGI will almost always be more intelligent than non-learning agent (assuming same history)
 - More knowledge can sometimes imply more intelligence
 - Learning is not part of the definition of intelligence
- AGI research has an issue: there's too many parts that might be considered intelligence, unclear what is really important

- Isolated research on different parts of intelligence (analogy independent of decision making) or (reinforcement learning independent of memory) or (discovery independent of ...) or (language independent of using it)
- Cognitive architecture considers all these components to build the “whole elephant”
- Natural selection is species improvement (not individual)
 - Not agent improvement, but cognitive architecture improvement
 - No direct transfer of learning by agent to next generation
- Analysis of chimpanzee utterances
 - No significant syntax or semantics
 - Semantics associated only with individual words
 - Koko “Who are you?”
 - * All answers contain “Koko”, but grammar is horrible despite being exposed to tons of syntactic structure
 - * May be able to understand but not generate syntactic structures
- Animals
 - Animals can’t do numbers, symbols, or language well
 - Animals do well with object perception, reasoning, physical properties, containers, etc.
 - Persistence, hypotheticals, gravity, circular motion
 - * Takes time (adults)
- Causality vs coincidence
 - Create underlying model of dynamics of the world
 - * Causal learning
 - Need to create causal models otherwise you just learn perceptual features and co-occurrence
 - * Deep learning systems
 - Seems spotting in chimpanzees
 - * Related to tool use
- Theory of mind
 - Chimpanzees probably can’t distinguish intentional behavior from accidental behavior
 - Model of others’ thought processes (desires, knowledge, emotions)
 - Can I create predictions from this?
- Teamwork (shared intentions)

- Mental time travel
 - Imagine, plan, prepare for the future
 - Squirrels store and recall food
 - Rats associate sickness with taste of food ate hours earlier
- Octopi learn too fast
 - Too much bias to effectively learn strategies
 - Only 1 in 10,000 hatchlings survive because of this
 - They can learn extremely specific things
 - Camouflage behavior
 - * You can learn something specific to the hardware
- Ladder of cognitive abilities
 - Asocial reasoning
 - * Solitary animals
 - * Innate and individual learning
 - Social reasoning
 - * Environment shaped by others
 - * Follow, cooperate, play, lead, warn, trick
 - Animal cultural reasoning
 - * Communicative teaching
 - Oral linguistic reasoning
 - * Evanescent
 - Literate reasoning
 - * Permeant rendering
 - * Copyright
 - Civilization-scale reasoning
- There's not just 1 option for AGI
 - Not even just 1 unified progression
- Hierarchy of cognitive capabilities across animals
 - Can start bottom-up or top-down or do both simultaneously
- What can we learn from animals about AGI?
 - Human-level AGI is probably impossible to start at the higher rungs of the cognitive ability ladder and fill in the lower level processes later

- Theory of mind
- How do we guarantee “good choices” for an AGI?
 - * We can look at natural selection and look at a higher level of the better performing capabilities

Journal

- Humans and animals share similar underlying components and structures (neurons, brain sections)
 - However, there are sets of components that are unavailable to animals. There is likely some innate difference in either the structure or plasticity of our brains. Is this due to theory of mind, or some innate prior?
 - I.e., we can rewire and optimize certain parts of our brains for things we want to focus or specialize on.
- Is it possible that we’re missing socialization in our AGI systems?
 - To go up the ladder of intelligence, we need basic understandings of trust and cooperation
 - After teaching 1 AGI, it can propagate the knowledge to others
 - We do not want to create AGI sociopaths which do not understand the meaning of empathy, trust, and cooperation
- Can we evaluate the progress of AGI using conceptual benchmarks?
 - E.g., does it understand object permanence, can it figure out what things it can interact with, does it transfer knowledge to other tasks, etc.
 - Might be a good idea to just create a gigantic checklist of components and go down the line from easiest and most important
- Use an algorithm to find the most memory and computational efficient architectures or technique for a given task

2022-01-16 – AGI Readings 4

Toward the quantification of cognition

- A brain is simply a finite state machine carrying out computational steps and a set of memory operations.

- If an AI makes a “mistake”, it’s merely a mistake to us humans since it doesn’t make sense to us. It’s just empiricism.
- We’re focusing on big data rather than focusing on open-ended problems
 - Tasks and models are basically just data memorization with slight generalization
 - * Failures and shortcomings are disregarded
- Evolution of brains
 - Overall brain size changes and evolves
 - Novel brain attributes are preestablished
 - Differences between brains of humans and other primates are very small
- Humans are different from other primates
 - Cultural learning, social interaction, language
- Some humans are special
 - Some are scientific geniuses
 - Information can be transmitted via recording and writing
 - Human with access to writing and recording isn’t the same as a human without storage
- Some human capabilities are unique
 - Human-level math
 - Mental time travel
 - Advanced cultural learning
 - Morality
 - Logical inference
 - Advanced social mindreading
 - Numerical abilities
 - Natural language
 - Theory of mind
 - Serial sequencing (ordering)
 - Working memory
- Human anterior cortex is larger than any primate by an order of magnitude
 - Computational power and memory size of humans is immense
- Nested stack automata and memory
- Convert experience into abilities
 - We need arduous training, resources, memory extensions to learn writing, logic, or math (complex skills)

- Why does the brain enable effortless language learning, but effortful learning of other tasks (writing, logic, math)?
- Why is this system absent in other animals?
- Humans could be Turing machines that lack the ability to exactly know their infinite tape memories (card matching game)
 - Is it possible that we compress and overwrite our memories via DNN or knowledge graphs
 - Episodic memory
- Brain size is not everything
 - Autistics with larger brains have language deficits
 - Children with 1 hemisphere removed before language acquisition display normal language expression and comprehension

Lecture

- Qualitative categories of human capacities
 - Ancestral
 - Core
 - Developmental
 - * Non-human versions of human capacities: antecedents
 - * Children have antecedent versions
 - * Communication, memory, social reasoning
 - * Some animals have primitive versions of seemingly human-only capabilities
 - Requires explicit training
 - Not very general
 - Not all members of species can learn ability
 - * Some skills animals have that humans don't have
 - Better memory, navigation
 - Perceptual
 - All-and-only abilities
 - * All humans do except for disability
 - * Only humans can do
 - * Learned effortlessly
 - * Language, counting, mental time travel, perspective tasking
 - * ***Is there a core that gives rise to all of these? Are they independent components? Are they all or nothing?***

- Expert
 - * Only human, only humans who deliberately train
 - * Writing, chess, programming, logic
- If we go from small animals to humans
 - Brain size predicts growth of specific sections of the brain
 - Neocortex, cerebellum, striatum, septum
 - * Relative sizes scale linearly to overall brain size
 - * Humans are better at memory of current situation and meta-reasoning
 - * Human brains are unexpectedly large for our bodies
 - Those later in evolution, these brains become bigger and bigger
 - No change in connectivity
- Automata and grammars (FSM + Memory)
 - Subregular -> regular, context-free, mildly context sensitive -> context sensitive -> recursively enumerable
 - Most of these memories must be infinite to actually do these things (or effectively infinite given a certain problem)
- Humans are more powerful with external memory aids
 - Human memory is not a stack or tape
 - Multiple long-term memories with different storage and retrieval properties
 - * Retrieval from semantic and episodic memory
 - * Bias towards recent memory
- 98% of human brain is cortical-subcortical loop circuitry
 - You don't just throw something at the brain and something comes back
 - Not just feedforward or local recurrent network
 - There's constant looping which allows us to recognize:
 - * Sequences of categories of sequences of ...
 - Sequences of clusters of sequences of clusters
 - **Look into transformers & heads, RNNs, etc.**
 - * **Allows abstraction and temporal relationships**
 - * Recognizing grammars, position-based categories, relational representations
- Migration of memory from hippocampus to neocortex
 - Episodic memory (hippocampus) doesn't do any generalizations

- * Is sufficient for simpler domains with less variation
- Neocortex does category/relational/semantic memory
 - * Necessary for complex domains where abstraction is necessary

Journal

- “Human minds are nested-stack automata, intrinsically computing indexed grammars. Our natural languages clearly reflect this;”
 - While this is likely true at the lower levels of computation, it is unclear how larger structures and areas of the brain interact.
- Why do humans lack exact memory?
 - We often lose connections to memories or add conflicting memory nodes.
 - Do we compress our memories?
 - * Could this have something to do with memory being episodic and experience based?
 - Is there some trade-off between learning capability and memory capacity?
 - Building something with unreliable memory forces us into different worlds.
- To achieve the effectively infinite memory needed for the discussed finite state machines, maybe compression is a key
- We should not assume that human-inspired cognitive or neuroscience components are feasible or the most practical for an AGI. While evolution and darwinism might suggest this, there has only been 1 species that has successfully achieved complex intelligence and human all-and-only capabilities. Most of evolution has focused on the ancestral, core, and developmental capabilities (innate, movement, perception, associations).
 - It may be unreasonable to assume that modeling neural circuitry is a scalable and practical approach for computers.
- Grammars and DNN structures
 - Transformers
 - * You have a sequence with associated features (vectors)
 - * For every point in the sequence
 - Query, key, value
 - Use query to pick another key which gets a value
 - Can do N queries (heads)
 - Allow for complex abstractions

2022-01-19 – AGI Readings 5

Foundations of Cognitive Science

- Symbols
 - Representational of other things
 - Examples
 - * Words in sentences, numbers in equations, atoms in formulas, directions in signs, objects in pictures, components in circuits
 - Symbols stand for something, the appearance of the token means something
 - Access and retrieval of symbol and the context of its structure
 - * Processing of that symbol and structure
 - Information theory (measurement of symbol in bits)
- Symbol systems
 - Memory contains symbol tokens
 - Memory, symbols, operations, interpretation, capacities
 - Knowledge is about its domain
 - * Describes a system whose behavior can be computed
 - How do symbols in symbol systems represent something external?
 - * Through perception and actions with the external world
 - Knowledge systems cannot be realized by symbol-level systems
 - * No finite device can realize an infinite set
 - * Need to attain good approximations
 - AI is dedicated to finding symbol-level mechanisms that closely approximate knowledge
- Architectures
 - Fixed structure that realizes a symbol system
 - * What about mutable architectures?
 - In humans, there are fixed structures and changing architectures
 - * Changing
 - Lifetime 10^9 seconds
 - Development 10^6 seconds
 - Skill acquisition 10^3 seconds

- Knowledge acquisition 10 seconds
- * Fixed
 - Performance 1 second
 - Temporary storage 10^{-1} seconds
 - Primitive actions 10^{-2} seconds
- Continuous plasticity means that architecture is the boundary separating structure from content
- Interrupt, protection, dynamic resource allocation,
- Intelligence
 - If a system uses all the knowledge it has, it is perfectly intelligent
 - If a system does not have some knowledge, failure to use it cannot be a failure of intelligence (failure of knowledge)
 - If a system has some knowledge and fails to use it, there is a failure of some internal ability (lack of intelligence)
 - Intelligence is relative to goals and relative to knowledge
 - * Know some commonsense things (addition, navigation)
- Search and problem spaces
 - For tasks that are difficult, the agent begins to search
 - * Assumes the agent is operating within a problem space
 - Problem search
 - * Search problem space described
 - Knowledge search
 - * Search memory of system for knowledge to guide the problem search
- Preparation vs. deliberation tradeoff
 - Store vs. compute tradeoff
 - How much time and storage to spend preparing vs. time spent analyzing and deliberating?

Human Cognitive Architecture

- Real time constraints on human cognition
 - Large degrees of freedom for constructing architectures of symbol systems
 - * Trying to guess what brain design evolution has opted for
 - * Evolution is dependent on local environments

- Human is a symbol system
 - Approximations of knowledge systems
 - Humans can emulate a universal machine, slowly, but eventually
 - * Spend time memorizing new states
 - * Humans invent new adaptation very fast
- Human architecture is a hierarchy of system levels
 - System levels are collections of components that are linked together and interact
 - Levels are abstractions, ignoring some of the things at the level beneath it
 - On the upper rungs of the hierarchy, size increases and speed decreases
- Time scale of human action
 - Evolutionary band
 - * Millions of years
 - Historical band
 - * Years, millennia
 - Social band
 - * Months, weeks, days
 - Rational band
 - * Hours, minutes
 - * Tasks
 - Cognitive band
 - * Seconds, milliseconds
 - * Unit tasks, operations, deliberate acts
 - Biological band
 - * Milliseconds and microseconds
 - * Neural circuits, neurons, organelles
- Biological band
 - Neurons spike or pulse
- Neural circuit level
 - Collection of neurons performing some function
 - Brain is a large collection of local circuits connected by distal cables
 - Activation (medium, continuous quantity)
 - * Summed, thresholded, transformed in local circuits

- * Some statistical average of neural pulses
 - Stimulus intensity is encoded as the number of pulses per second
- Real-time constraint on cognition
 - Sentence is uttered in 1 second, reaction to it takes about 1 second (comprehension, decision, implementation, movement)
 - A system can't play 10 minute chess if it takes 1 minute to engage its cognitive behavior
 - There are only about 100 10ms periods to accomplish all computation
 - * About only 100 operation times
 - * Luckily, we have parallel systems and GPUs
 - * Lowest level of deliberation must occur at about 100 ms
 - Systems must work on the order of 1 second
 - * Inference must be very fast
 - * Spend downtime training and learning
 - * Reading upside down is much slower
 - Response time is decreased by preparation
- Intendedly Rational Band
 - Why are there not more and more infinite levels of the cognitive band
 - * As the system has more time, it can find better and more efficient solutions. However, this is capped by knowledge.
 - * As time goes on, the system becomes harder to understand
 - For reasoning and rationality, the agent does whatever its knowledge permits in order to attain its goal
 - * It may go outside of a context
 - Can groups be described as knowledge-level systems?
 - * As a single body of knowledge and set of goals?
 - Humans communicate knowledge too slowly to one another

Lecture

- Allen Newell
 - First AI program, multitask AI program
 - Turing award
 - CMU
 - Helped start HCI, cognitive science society, AAAI

- John Laird's thesis advisor
- Knowledge search vs problem search
 - Knowledge controls and informs search
 - * Fast and fixed mechanisms
 - * Search is over existing organized knowledge
 - * No awareness of processing
 - * Deliberately controlled knowledge search
 - * Kahneman's System 1
 - 95% of what we do is system 1 according to Kahneman
 - System 1 implements system 2
 - Problem search is serial
 - * Generative, combinatorial
 - Combinatorial is extreme case (knowledge usually gives intuition)
 - * Controlled by knowledge
 - * Improves with experience
 - * Aware of processing
 - * Kahneman's System 2
- Preparation helps convert problem search to knowledge search
 - Difficult problems require constructing problem space (which humans are bad at)
- Remote associates test
 - Given 3 words, come up with a 4th which is associated with all three
 - * 7-15 seconds
 - * Swiss, cake, cottage = cheese
 - * Dew, comb, bee = honey
 - * Man, glue, star = super
 - Difficult, man -> person -> etc.
 - An example of problem search controlling knowledge search
 - * Retrieval problem
 - * If you get the answer, you're like oh yeah of course
 - SWOWEN and HBC perform well
 - * Google Books doesn't do well
- Symbols
 - Can have partial information with complete structure

- * President of USA
 - Can use this without using Biden or etc.
- * Must use in context with other things
- Efflorescence of adaptation
 - Humans just go around creating opportunities to build different response functions
 - * Strong intrinsic motivation for discovery
 - * Can quickly learn new tasks
 - * Do new things that are different than before
- Kahneman's System 0
 - Neuron, organelle, biological
- 10 seconds
 - Complex reasoning
 - Analogy
 - Planning
 - Meta reasoning
 - Theory of mind
- 1 second
 - Simple reasoning
 - Mental imagery access
 - Language processing
- 100 milliseconds
 - Reactive decisions
 - Skilled behavior
 - Primitive internal actions
 - Access long-term memories
- The above here can be done in many instructions
 - Not just do all in 1 instruction
- Promiscuous intermixing of cognitive capabilities
 - Not module by module
 - Parallelism
 - Synergies

- * Using some component in another one
- The more you do something, the faster you get at it
- Which of these components do you need to be aware of to be able to do higher level cognition?
 - Obviously don't need neuron information to do meta-reasoning
 - Maybe need neural circuitry or higher level components
 - Some behaviors can be predicted solely by the structure of that level (strong levels)
 - * When things go wrong or we see the lower levels pop out, that's a weak level
- Stability is more important than efficiency
 - Like assembly code for computers
 - Like neurons or neural circuitry
- Seibel task, cigar making
 - Perfect power law scale
- Hypothesis in HCI and cognitive modeling: The deliberate processing is the fastest and strongest level
 - 50ms for basic operation, 100ms for memory, 70ms for finger movement, 100ms for basic memory

Discussion questions

- Computations not equivalent to search?
 - Generative computations with no particular objective
- Ability to have goals across such long timescales?
 - Emotions, intrinsic motivation for these long-term goals?
 - * Immediate reward or long-term reinforcement learning?
 - Social expectations? Cultural?
 - * Institutional, familial,
 - * Parenting
 - Cultural or biological evolution?
 - * Related to age, experience, development
 - * Parenting
 - Maslows hierarchy of needs

- * Don't need to prioritize
 - * Top-level is finding meaning
- Neocortex (larger prefrontal cortex)
 - * Memory also with long-term goal
- Are animals aware that they have a long-term goal?
 - * Do they contemplate on homes, seasonal migration?
 - * Mental time-travel, theory of mind
- How do hybrid neuro-symbolic AI systems collide with Newell's analysis of levels
 - Implications of having strong deliberate and neural circuit levels for AGI?
 - * Neural firing over time
 - Have something that interacts at a meta-level with neural circuits, neurons, and higher level structures
 - Abstraction bands within abstraction bands

Journal

- Assuming that humans are symbol systems, we must have very vast memories and very accurate and efficient lookup systems
- "There is no way for a social group to assemble all the information relevant to a given goal, much less integrate it."
 - Not sure this is true given the context of human history, trust, and cooperation. Albeit slower, there is still a gathering of knowledge, consolidation of goals, and group reasoning to achieve tasks.

2022-01-22 – AGI Readings 6

Objects

- Cohesion, continuity, contact of objects
- Infants are only able to represent a small number of objects at a time
- When attention resources are stretched, the finer distinctions fail to guide object representations
 - Such as foods, tools, etc.

Agents

- Infants represent agents' actions as directed towards goals
- Agents need not have faces and eyes
 - Can recognize intent, goal-directed actions
- Goal directedness, efficiency, contingency, reciprocity, gaze direction

Numbers

- Number representations are imprecise, growing more imprecise with increasing cardinal value
 - Number representations apply to diverse entities encountered through multiple sensory modalities
 - Number representations can be compared and combined by operations
- Infants can discriminate between large numbers of objects, actions, and sounds
 - They use a ratio limit on precision
 - Approximations

Geometry

- Distances, angles, relations
- Doesn't handle color, odor, etc.
- Children suck at geometry
 - Adults use landmark objects

Us vs. Them

- Identify and reason about potential social partners and social group members
- Race, ethnicity, nationality, religion, arbitrary assignment
- It's us or them
- Biases for same race
- Group membership from language
- Is this really a necessary or useful feature to have in modern society or for AGI?

Other

- Symbol systems, language, maps, reasoning about physical/biological phenomena, cognitive skills, skill acquisition

Lecture

- Katherine Kinzler
 - Professor of Psychology at UC, chair at Cornell
- Elizabeth Spelke
 - Cognitive neuroscience, specifically young children
- Systems
 - Objects, motion, mechanical interactions
 - Agents and goal-directed actions
 - * Action without contact
 - * Self-control and goals
 - * Children take non-wasteful short paths
 - Sets and numerical relationships
 - Places and their geometric relationships
 - Social partners
 - Each system has entities (rules), they support inferences about their interrelationships and behavior
 - Concepts vs architectures needed to process these concepts
 - * Do architectures enable these concepts? Do these concepts imply specific architectures?
 - Are other systems subsumed by these core systems?
 - * Maybe these core systems are so fundamental to children, and that indicates everything else builds on this? (innate)
 - Other core systems can be prioritized or learned (chess)
 - Propensity to group/cluster things together
 - * Need enough causal models on these core systems
- Objects
 - Persistence

- Children can only focus on things right in front of them (attention mechanism)
 - * Could be adapted
 - * Embodiment changes, environment changes
- Motion
 - * Cohesion, continuity, contact
- Rigidity (slinky)
- Non-cohesive objects (sand, fluids, smoke, snow, knowledge, trash)
- Essentialism (Susan Gelman)
 - * Women, racial groups, dinosaurs, original Picasso artwork (an underlying reality/nature that one cannot observe directly)
 - * This essence gives objects their identity
 - * Responsible for similarities that categories share
 - * Essentialism is a reasoning heuristic (bias) available to both children and adults
 - * 2 levels
 - Observable reality
 - Level of explanation and cause
 - Philosophy (human cloning, replaceable parts on boats)
 - * Hypotheses
 - Half the US population reject evolution, we are in essence humans
 - 1/3 believed they would take on new personalities or characteristics when receiving a heart transplant
 - People place value on authentic objects instead of exact copies (“Original is better, NFTs”)
 - Overgeneralization
 - Explains how people are screwed up...
 - AGI have to understand these human aspects
- Core systems vs Newell’s knowledge and process
 - Content vs knowledge/process
 - Core systems are innate, so you must be able to process them very quickly
 - Core systems are focused on the representational systems
 - * Are the core systems the smallest possible units (atoms)?
 - * Are agents actions?
 - Are core systems sort of a system 2.5 for Kahneman?
- Final project: build an AGI system
 - How to incorporate core systems into an AGI system?

- * How are new skills, concepts built on these core foundations?
- * Innate perceptual bias?
 - DNN bias towards features (colors, edges, shapes, motion, etc.)
- * Innate perceptual categories?
 - Classify things into categories of core systems
- * Innate process models of dynamics between different types of entities?
- * Something more basic than core systems:
 - Thing
 - Fragment of space and time
- * Does this imply a curriculum?
- * Specifically for reasoning
 - ***Building a model of our world***
 - Not for meta-reasoning, perception,

Journal

- Core systems
 - Thing
 - * Each thing has a set of properties and interaction/relations with other things. These things can be represented by symbols.
 - * Objects
 - Physical interactions with other objects (containers, momentum)
 - Color, shape, usage, etc.
 - * Agents
 - Theory of mind
 - People, cats, dogs, (plants?)
 - Intentions, goals, actions, etc.
 - * Places
 - Setting in which objects and agents reside in.
 - Location, time, size, cleanliness, etc.
 - Core systems can be used to represent episodic and semantic memory easier, by having some structure for each thing/concept.
- Essentialism
 - Can boil down the essential properties of each core system to create a simplified and generalizable representation of a thing.

2022-01-26 – AGI Readings 7

Embodiment of concepts

- Background on concepts
 - Enable us to have attitudes, beliefs, and desires about objects/concepts
- What does it take to learn a concept?
 - Rely on senses and perception
 - Use relevant background knowledge and infer
- How are concepts represented?
 - Abstract list of properties or relations
 - “Cow” = body of information related, used to categorize, analogies, and language
 - * How can a simple set of properties allow for categorization, inductive/deductive reasoning, and language understanding
 - Prototypes (statistical) vs exemplars (concrete) vs theories (descriptive)
 - Concepts themselves are like arbitrary/abstract words or symbols. The actual meaning comes from its interactions with other things.
 - Connected to core systems
 - * These provide initial expectations/representations which are updated
 - Conceptual representations are flexible
 - Predictive models are updated whenever there are errors
- Bottom-up vs. top-down
 - Seeing a cow activates anticipation and inferences about the behaviors and senses associated with cows
 - Prior models of cows generate top-down predictions to sensory systems

Intelligence without reason

- Traditional AI is top-down
 - Tackles AI using thinking and reasoning
- Bottom-up AI
 - Physical modeling, robots, tasks

- Robots
 - Tried to sense/model a 2d or 3d world, plan things, and make actions
 - Way too slow in sensing/modeling, not enough computation in planning/acting
 - Most of what we do on a daily basis is just routine
 - Must build them in real dynamic environments and have them operate in real time
 - Situated in the world (not abstract), embodiment and exist in the world, intelligence from sensing and processing, emergence of intelligence from interactions
- Computers
 - Slow and steady progress in research, occasionally big discoveries that grow fast and branch off into their own thing
 - Biological systems are massively parallel, computers are not
 - Minimax search, optimizations, trees,
 - Complex semantics into blocks
 - Backpropagation, useful but slow
- Biology
 - Should information be stored independent of the way it was collected/used?
 - * There have to be some amount of links
 - Neurons and circuits are the foundation
- Ideas
 - Blurring between world and AI knowledge
 - Bounded rationality
 - Use the actual world as the best model
 - Need an ongoing participation and perception for there to be *meaning*
 - Intelligence emerges as a result of interactions and dynamics of the world
- Thoughts
 - The author argued for situatedness, embodiment, reactive architectures
 - * Not for reasoning systems, manipulable representations, symbols
 - * For decentralized computation
 - Computation is done asynchronously with a bunch of active components. Messages are communicated between senders and receivers and are context-dependent
 - There is no central model maintained over the world, it is distributed
 - There is layering of abstract complexity over existing networks
 - * Think the different cognitive levels and limiters

- No hierarchical arrangement, everything in parallel
- Nodes freely allocate themselves
- Learning
 - Trade-off between innate genes and learning
 - 4 classes of things to be learned
 - * Representations of the world
 - * Aspects of instances of sensors and actuators
 - * Interactions of individual behaviors (relations)
 - * New behavioral (relation) models
- Important features
 - Convergence, synthesis, complexity, learning, coherence, relevance, adequacy, representation, learning, emergence, communication, cooperation, individuality,
- It is unclear whether this will all lead to theory of mind.

Grounded cognition

- Cognition is grounded in real things/experiences
- How do metaphors work?
 - Love can be eating
 - Happy is up and sad is down
- Memory theories focus too much on passive storage and not enough on situated action and retrieval
- Stimulus leaves memories in the modal areas that encoded it
- Physical simulations are piecemeal and sketchy, no holistic and detailed
- People conceptualize time using spatial trajectory (horizontal or vertical)
 - However one can also conceptualize time using episodic memory and numerical computations (10 minutes \times 6)
- Causal learning, statistical learning, and symbols
- Mirror neuron systems and social cognition
- Is the brain a single representational system?
 - How does the brain represent abstract concepts?
 - * Heavily dependent on context
 - Is it hierarchical in nature?

Lecture

- Blind men and the elephant
 - Want to develop theories where there is a single “big” idea for everything
 - * Symbol systems, dynamic systems, DNNs, reinforcement learning, analogy, language, grounded cognition/simulation
 - This happens because people focus on specific types of tasks
 - * We use representations appropriate for the task
 - * Logic, math, abstract thought = symbols
 - * Language, words, sentences = language stuff
 - * Puzzles, games, lots of uncertainty = search spaces
 - * Perception interpretation, categorization
 - * Sequential decisions without memory but with reward
 - * Navigation, spatial reasoning = grounded cognition
 - We need to integrate all these things
- Larry Barsalou
 - PhD Stanford, cognitive psychology
 - Faculty at Emory, GTech, Mich, Chicago, etc.
 - Theme is that conceptual system is grounded in multimodal simulation
- Key concepts
 - Amodal symbols
 - * Independent of sensory information
 - * Pointers to modality-specific information
 - Like word2vec or memory pointers
 - * Unifies the cluster of multimodal representations
 - Modal symbols
 - * Includes information from how entities were sensed
 - * Multiple modalities (touch/smell/written/spoken)
 - * Maybe when you hear about Mark Hamill, it’s associated with sound, but when you see Mark Hamill, it’s visually associated
 - * What operations are performed on modal symbols?
 - Mental imagery
 - * Classic examples of modality specific representations and reasoning

- Imagine you're on the northeast side of an island in the shape of the letter that comes after "B". How much of the island must you traverse to get to the southeast side by walking?
- Being able to do this without creating a mental image
- People who do not have good mental imagery do not think it exists...
- Matthew?
- Simulation is the reenactment of perceptual, motor, and introspective states acquired during experiences with the world, body, and mind
 - * Is any processing or active representation equate to simulation?
 - * More like imagination and the unique combinations created by this
- Grounded Cognition
 - Human visual experience is not continuous
 - * Saccades happen ~3 times/second
 - * We are effectively blind
 - Aphantasia
 - * Doesn't impair creativity, but people lack ability to create mental imagery
 - We need to define both data structures and processes
 - * Semantic networks (people would talk about structures and links, but how would you reason or provide algorithms for them?)
- Knowledge and conceptual processing
 - They tested their theories on perceptual features (size)
 - But if they tested it on semantic relations like remote associate tests: Bee, comb, dew, etc.
 - Be skeptical of generality claims
- Takeaways
 - Doesn't seem to prove absence of amodal symbols and reasoning
 - What can multimodal representations and reasoning do for tasks?
 - Mirror neurons are magical???
 - * Theory of mind for activities and mapping onto other people and learning from imitation
 - We share neural activations for similar things
- Grandmother neuron theory
 - Representations are distributed, it's not like the neuron of your grandma dies and you lose memories

- MRIs suggest it's possible to give you a picture of a sailboat and read the fMRI machine to know what you're looking at
 - * Activation of brain is distinctive and indicates classification
- There is a grandmother distribution of activation that is close to a "grandma"
 - * We experience grandmas similarly, so there's some shared ground ans sensing

Discussion questions

- What is a precise definition of simulation and contrast with other types of processing?
 - Some sort of dimensionality reduction of the environment
 - * Only saves the essentials
 - * How do we derive or focus attention the essential information
 - Construct an amodal representation using the essential modal information
 - Daydreaming
 - * Recreating the world around us and environment in order to generate and imagine new scenarios.
 - People replay the ABC song if asked about the 13th letter of the alphabet
- How do topics covered map onto core concepts?
 - Multimodal representations underlie core systems as a foundation
- How do topics covered map onto Newell's levels?
- What are the main takeaways for AGI and how do they influence an AGI design?
 - AGI need to define the essentials and attention mechanisms
 - Language fetches amodal representations and uses modal representations together
 - * Top-down
 - * Modal representations and perception to language is bottom-up representation
 - Perceptual inferences and linguistic inferences
 - * Using contextual information, using simulation
 - Need to figure out which modality is the easiest to perform a calculation with
 - * Modules available for reasoning with symbol, perceptual senses, picks the best one

Journal

- Rather than having innate priors of concepts, do we formulate them ourselves using what we perceive and find patterns in the world/environment around us? For example, why is it that the

first few concepts children learn are shapes and colors? Is this due to our curriculum or is this the extent of patterns that they are able to grasp? How do more complex concepts build on top of existing ones? Is there some ordered hierarchy of abstractions/concepts?

- Perhaps the concepts build upon one another, and we are able to imagine/generate new concepts using the foundational concepts and properties.
- Even for abstract concepts that seemingly have no modality, e.g., hope can be represented by a warm/positive feeling for the future.
- “If concepts were solely based on perceptual experience and passive associative learning, it remains unclear how children learn to distinguish between concepts (e.g., living vs. non-living) that are based on perceptually similar experiences (e.g., a living bird vs. a stuffed bird).”
 - I can see both sides of this argument. A child can perceive a living bird moving, flying, and chirping and over time passively associate these qualities with life. One could conceptually differentiate between the two using context clues from the environment as well (presence of trees, other birds, etc.) However, if a realistic stuffed bird were placed on a tree, one may use predictive modeling to discern finer-grained details via reinforcement learning. “My prediction was wrong. It was actually a stuffed bird. These are the minor differences to keep track of next time.”
- Subversion of expectations can provoke a strong emotional response (surprise) and inform future predictions.
- Is language just a convenient tool for fetching concepts and relations?

2022-01-28 – AGI Readings 8

The computational origin of reasoning

- Intermediate bridging below symbols but above neurons/network nodes
- It is possible to compute without notions of logic and numbers
- Language of thought (LOT)
 - $\text{lift}(x, y) = \text{CAUSE}(x, \text{GO}(y, \text{UP}))$
 - * Shop-lifted? Ride lift?
 - Has to be the concept lift, not the word lift
 - Conceptual role semantics (CRS)
 - * Meaning derived from relationships with other symbols
 - * Imagine taking a complicated physics course where you know no terminology

- You figure it out via relations and applications to the real world
- * 11010 is a number in binary, but to an alien it might mean a different thing depending on its use
- * Isomorphisms
 - Symbol relations should mirror real-world relations
 - Can construct a representation of pretty much anything
 - How do symbols get their meaning?
 - Architectures for symbols seem unbiological
 - The computations between symbols matters more than the symbols themselves
 - Need a strong AI that is able to understand these things
- Combinatory logic
 - Mathematical reduction
 - They create an exhaustive program that maps symbols into combinatorics
- Cognitive domains
 - Need to be able to generalize these relations
 - Can create and represent structures like trees, lists, recursion, repetition, identity, if/else
- LCL theories
 - Handle abstractions using variables
 - Are dynamical, emergent, parallelizable
 - Supports generalization using property induction, deduction and simulation, and learning
 - * Bayesian learning
- Connectionist architecture

Commonsense Reasoning and Commonsense Knowledge in AI

- Super slow progress
- Figure out ambiguities in text using context
 - Pronouns
- In computer vision, ability to infer existence of things that are not directly visible
- Successes in common sense reasoning:
 - Taxonomic reasoning
 - * Collections of categories, individuals, properties, and their relations
 - * Linked using subcategory, instance, property of, and cancels

- Temporal reasoning
 - * Pretty much solved, except integrating with natural language is difficult
 - Context-dependent interpretations, natural language expressions are complex
- Action and change
 - * Solved if:
 - Events are atomic, deterministic
 - Consider only states before and after events
 - Single actor
 - Perfect knowledge
 - * Unsure about:
 - Continuous domains, simultaneous events, probabilistic events, imperfect knowledge, decision theory
- Qualitative reasoning
 - * Trade-offs and causal effects
 - Prices go up, number sold go down
 - Temperature of gas goes up, pressure goes up
 - * Physical motion is difficult
- Challenges
 - Modeling a teacher thinking about what their students don't understand is difficult
 - Coming to conclusions that are reasonable does not mean they are correct
 - * Unreliable data, rules which conclusions are likely but not certain
 - Some examples are very frequent, but some have a long tail of many infrequent examples
 - * "of the year" vs "moldy blueberry soda"
 - Discerning the proper level of abstraction is difficult
 - * How broad/narrow should rules be formulated to be?
 - All sharp objects stuck in other objects create holes?
 - Not in sand or water...
 - It's context-dependent, no?
- Commonsense approaches/techniques
 - Web mining
 - * NELL, KnowItAll
 - Mathematical
 - * Situation calculus, region connection calculus, qualitative process theory

- Informal
 - * Scripts, frames, case-based reasoning
- Large-scale
 - * CYC
- Crowd sourcing
 - * ConceptNet
- Different ranges of reasoning modes, breadth, application-dependence, etc.
- Alice, bob, carol are playing together. Someone says “at least one of you has mud on your forehead”. Alice and Bob say “I don’t know”. Carol says “My head is muddy”.
 - Basic reasoning about this generates several potential explanations: Carol feels her forehead is muddy, Carol fell down in the mud, Carol assumes her head is muddy since Alice and Bob weren’t sure, Carol is lying.

Lecture

- Tesla’s humanoid robot (Optimus)
- Ernest Davis and Gary Marcus
 - Critics of over-enthusiasm of AGI
- Common sense vs paper
 - Paper focuses on knowledge and reasoning in general
 - Godfather example is not common sense
 - Emphasizes need for formalization
 - Is this an AI-complete problem?
 - * To solve this problem, you need to solve all of AI
 - What even is common sense, just a shorthand for something? There’s no special sauce to common sense
 - * Don’t touch the stove when it’s red hot
 - * Are there lots of special sauces or no special sauce at all?
 - * Common sense is the removal of all specialized knowledge
 - We won’t have reliable AI until we solve common sense
 - Argues against certain approaches (statistical correlation)
 - Big question:
 - * How is common sense knowledge and reasoning different from other types of knowledge and reasoning?

- How does an internal combustion engine works?
- What is mRNA?
- What city is UMich in?
- One definition of common sense
 - Knowledge that all humans have, unspoken, unwritten - we take it for granted.
 - * “Animals don’t drive cars, my mother is older than me”
 - * Culturally specific or discipline specific
 - * *Maybe it’s something you’ve seen so much that it becomes a system 1 thing?*
 - Coming back to the AGI -> specialized AI design
 - Emphasis on general knowledge:
 - * “Ice floats” is kind of specific
 - * 1kg of iron is heavier than 1kg of cotton
 - Incorrect common-sense? System 1! Usually pretty accurate, but after thinking more about it, it’s wrong.
 - Sometimes it’s not system 1 since you do knowledge retrieval and complex combinations
- Another definition
 - Human-like ability to make presumptions, conclusions, reasoning about the type and essence of ordinary situations humans encounter every day
 - * Is core knowledge/systems common sense reasoning or a small subset of it?
- Laird’s thoughts
 - Common sense knowledge is stored in long-term memories
 - * Was experienced or derived in the past
 - Common sense reasoning involves
 - * Applying common sense knowledge to a specific situation
 - * Deriving simple entailments of existing common sense knowledge
 - * Commonsense seems to have do with contextualizing problem-specific knowledge within border knowledge and pertain to the operations on the border of these 2 spaces
 - 80-20 rule
 - * Maybe it’s more the 95-5 rule for AI?
 - Unusual situations where you can’t rely on expertise: relies on general knowledge reasoning
- CYC: Knowledge base with micro theories

- Get to the edge of what things have been trained on, then they become brittle
 - * Robustness in ML
- Goal of Cyc was to have the knowledge needed to understand entries in encyclopedias
 - * Accumulation of knowledge is exponential, has not reached the exponential capacity yet
- COMET, ConceptNet will be discussed on Monday
- What things fall under common sense?
 - Taxonomic reasoning, temporal reasoning, action and change, qualitative reasoning
- Biggest challenges for common sense:
 - Seemingly endless scope
 - * No boundaries, long tail of infrequent categories/knowledge/...
 - * Maybe common sense isn't built up using categories of "things", but rather a set of "things" that are stored or obtained in a certain way (fast)
 - Because we always keep obtaining more and more knowledge (7yo common sense is different from adult common sense)
 - Lack of formal definition of what it is
 - Difficulty in evaluating whether AI systems have it or fake it
 - Lack of learning approaches that extract semantic information
- Winograd challenge
 - Systems don't have to develop a semantic representation of what's actually going on in the sentence
 - Looking for co-occurrences (biases)
 - Logic problem not language problem
 - * The trophy doesn't fit into the brown suitcase because it's too large
 - The trophy is too large
- Winograde
 - 44k problems
 - Systematic bias reductions using an algorithm that detects co-occurrences in language
 - SOTA models have lower accuracy (59.4%-79.1%) compared to humans (94%)
 - Availability bias
 - * <https://www.google.com/search?q=availability+bias&client=firefox-b-1-d&source=lnms&tbn=isch&sa=>

Discussion questions

- How is commonsense knowledge and reasoning different in kind from other types of knowledge and reasoning?
 - Maybe it's simple, frequent things that are processed using system 1 techniques.
 - * Regularization, needs fewer data samples to generalize
 - Morals and ethics
 - Common sense knowledge is the knowledge a speaker/writer would assume a listener/reader already has
 - * Common ground?
 - * Could extend to domain-specific knowledge
- What are strengths and weaknesses of different approaches to it?
 - Cognitive dissonance, common sense reasoning is just the 1st reason that comes to mind
- Once we really understand general learning, will commonsense K&R just fall out? Why or why not?
 - You can acquire knowledge through experience or by someone telling you
 - Not innate, more common knowledge acquired
 - * Multimodal also
 - Learning influences common knowledge
- Is commonsense K&R really needed for robust, task-specific AI
 - Probably not... Long-tail
 - Rare situations, infrequent
 - * Large state space makes commonsense relevant
 - Small number of situations that happen very often
 - * We can generalize this knowledge and reasoning to the rare situations (commonsense language-knowledge graphs are not a good mapping to human knowledge)
 - Robot experiencing the world in many modalities
 - Common sense will help, but is not necessary
- What would be better types of tests for commonsense K&R
 - No tests really seem to fill this.
 - It's an ability to learn all the commonalities of some experience.

Journal

- How do we automatically construct isomorphisms from the real-world?
 - We still need a way to map meaning to symbols via some sort of “environment modeling”. Can this be done using core systems and memory retrieval?
- How does learning using concepts work if they require prior knowledge of more foundational concepts?
 - There must be some system in place to begin picking up novel concepts via basic learning methods.
- Is commonsense reasoning more just simple/quick logical operations performed on existing knowledge in order to generalize rules to other situations/contexts?
 - Perhaps commonsense reasoning and knowledge can be seen as a way of warm-starting the learning process (not necessary, optional and helpful for learning/training speed)
- “of the year” vs “moldy blueberry soda”
 - Reminds me of GPT-3 and DALL-E (generating images from obscure captions)
 - * “Avocado chair”
 - * Possible reasons why this works very well:
 - Multi-modal representations of concepts
 - Very large network architectures (lots of “space” to fill up with knowledge)
 - Very large amounts of data (lots of experiences to fill up the space with)
- Perhaps common sense knowledge is not something to be seen as a component or part of the AGI architecture, but something to be used as part of the learning process or curriculum for an AGI
 - Particularly useful for mapping language to concepts and for creating a system to categorize and add properties to objects/agents/places
 - * Then you can create more generalizations after having some modal data and context
 - Maybe this is why transformers work particularly well. Unfortunately, having multiple or 1 gigantic transformer is not realistic for run-time or memory.

2022-02-02 – AGI Readings 9

Understanding

- It’s possible for something to live/survive without needing to understand the world

- Virus, plants, etc.
- Chaos and unpredictability are a reality
 - We want to create theories that are as accurate and precise as possible
- Association learning with strong vs weak associations
 - Must connect classes of stimuli and classes of responses
- Rats don't remember *exactly* how they walk through mazes, their actions slightly differ
- Knowledge is more than just something that holds information
 - The sum/whole of the parts of information must be greater than just the information
 - * Relations, retrieval, applying information
- Abstract thinking
 - Variables, facts, rules, inference
 - For all, there exists, some, etc.
- Chinese room - simulate without understanding
 - DNNs
- There is no upper bound on understanding
 - Infinite degrees of depth/layers of abstraction
 - Hierarchical chunking
- Case studies
 - Someone who's not great at cooking and understanding food science can still follow a recipe easily
 - * Infer that higher temperature means shorter cooking time
 - * Inability to creatively combine ingredients using knowledge of their interactions
 - There's a difference between no understanding and some understanding
 - * Also a difference between some understanding (cookbook) and expert understanding

What does it mean for AI to understand?

- Language models have reached 90%+ on Winograd benchmarks (SuperGLUE)
- DNNs use statistical shortcuts because it's "optimal"

Dark, beyond deep: A paradigm shift to cognitive AI with human-like common sense

- Previous/current AI for computer vision
 - DNNs and big data for small tasks
 - * Needs to be small data for big tasks!
- Reason about unseen things: human common sense
 - Object permanence
 - World simulation/recreation
 - These are “dark” areas
- Task-driven vision rather than data-driven
 - Example tasks
 - * Object detection/recognition, object manipulation, task planning, SLAM
 - Don’t need precise 3D reconstruction
 - * Grid cells, mechanism for cognitive representation of Euclidean space, seen in rats, bats, etc.
 - Fluent and perceived causality
 - * Transient state of an object that is time-variant
 - Cup being empty or filled, door locked, car blinking signals, telephone ringing
 - * Attributes are permanent
 - Functions
 - * Human actions
 - Intentions and goals
 - * Intent could be the transient state of agents
 - Thirsty, emotions, etc.
 - Utility functions for rational agents
 - * Markov decision processes
 - Bidirectional inference
 - * Combine top-down inference with abstract knowledge and bottom-up inference based on visual patterns
- Causal perception and reasoning
 - Can we see causality from vision like with color or depth?
 - A launches B (2 balls interacting)

- * Even if person is told they're just pixels, we still perceive it as launching
- * If there's a small temporal gap between A stopping and B moving, the illusion breaks down
- Causality is clearly important and likely requires some temporal backtracking.
- Create virtual escape rooms
- Statistical causal learning with randomized/nonrandomized studies
 - * Powerful if real-world interventions/actions are possible
- Sort of like a state machine or knowledge graph for fluents, causing actions, and actions
- Intuitive physics
 - We essentially have a mental physics engine
 - Stability and safety in scene understanding
 - * (Gravity)
 - Physical relationships in 3D scenes where objects support, attach, or hang from each other
 - Someone integrated physics engines with deep learning to predict future of static scenes
- Functionality and affordance
 - Interactable objects like switches, buttons, knobs, hooks, caps, handles
 - * Eventually we learn what shapes permit different interactions
 - Containers and tools
 - * Should figure out where/how to hold things and what motion to use it
 - * Will objects contain more objects?
 - Representing scenes with layouts, categories, activities, functional objects, attributes, etc.
- Intent/Agency
 - Represent a future goal state and try to achieve this
 - * Ability to plan by picking the most rational actions
 - Theory of mind attributes mental states to oneself and others
 - Infants segment continuous behaviors into goal-directed acts
 - * Perceive intentional relationships among environments, actions, underlying intent
 - * Infants can imitate actions in a rational and efficient way
 - Humans devote limited time and resources only to those actions that change the world according to their intentions and desires
 - * Achieve this by maximizing their utility while minimizing their costs, given their beliefs about the world
 - Action-effect association, simulation procedures, teleological reasoning

- * Actions with intent rely on simulating actions and mapping it into our own internal representations
- * Action-effect association plays a role in quick online intent prediction
- * Social learning is done via teleological action-interpretational system
 - Thinking about the purpose of something
- * These mechanisms complement each other
- Learning utility and choices
 - Utilitarianism
 - * Use some utility function
- “Essential AI ingredients”
 - Physically realistic VR/MR platform for big tasks
 - * Discretizing physics
 - * Eulerian grid-based approaches
 - * Lagrangian mesh-based methods
 - FEM
 - * Lagrangian mesh-free methods
 - SPH, RKPM
 - * Hybrid Lagrangian-Eulerian methods
 - ALE, MPM
 - Social systems (language, communication, morality)
 - * Communicate using a rule-based system? Markov decision processes and Q-learning?
 - Intelligence tests (IQ tests, reasoning tests)
 - * Analogies require understanding causes and effects
 - * Contrast effect
 - * Transfer learning and problem solving
 - * Number sense

Lecture

- Multi-university research institute (MURI)
 - Often funded by Department of Defense
 - * Office of Naval Research, Army, etc.
 - Vision/idea is to be very inter-disciplinary and have a big exchange of ideas

- * Often times, the money just goes to a few graduate students' paychecks
- Common sense knowledge and reasoning
 - Innate (Implicit)
 - * Geometry and physics integrated into sensors and motors
 - * (Scene understanding)
 - Individual experiences (Implicit)
 - * Functionality, affordances, causality, intent
 - * (Uses of a hammer)
 - Indirect experience: historical media exposure (Explicit)
 - * Cultural knowledge, general world knowledge, science, arts, history, schooling
 - * (Earth is a sphere, WWII)
 - Indirect experience: current media exposure (Explicit)
 - * Recent events, recent history
 - * (COVID, Olympics)
 - Essentially a dark matter, not super useful yet, but categorizes an area of capabilities that AI currently do not have
- Common sense and AGI
 - No single type of commonsense knowledge and reasoning
 - * Heterogeneity of knowledge
 - * Not a single CKR module
 - CKR integrates into agent processing in different ways
 - * Intertwining of low-level processing and representation
 - Take advantage of latent structure in domains
 - * “Small data for big tasks”
 - * Vision processing (latent structure not in pixels)
 - * Physics, causality
 - Sensory processing is general - not pure trained for single task
 - * No pure bottom-up from pixels to internal 3D model
 - * Incorporates general constraints from physics, affordances
 - Can these ideas be generalized to other types of processing (non-sensor, non-motor)?
 - * Reasoning, abstraction, induction/deduction/reduction
- Causality
 - Perceptual innateness

- * (Launching)
 - Complex aspects require reasoning
 - * (Determining the source of COVID)
 - Deep RL can't solve it all
 - * The AlphaGo, Alphaetc, implicitly encode causal relationships into their architectures
- Identifiability problem in Cog Sci
 - A system that produces the same behavior as humans for some task does not mean it uses the same underlying processing
 - * Predict tower collapse in brain's physics engine
 - * Raven's Matrices performance

Discussion

- How do different levels of CKR get incorporated into an intelligent agent's activities? Develop some categories and examples.
 - How do we use CKR in our daily lives?
- If CKR is not a meaningful categorization, what are potential alternatives that would be useful in AGI development?
 - Current DNNs don't have CKR, what is a sharper category of things they should have?
 - Connection to core knowledge/systems?
 - Connection to Newell's high/low levels of computation?
 - Defining CKR is NP-hard
 - (4) In general, commonsense knowledge & reasoning isn't a very well defined category, defining the subcategories would be better. Tried to differentiate between high-level vs low-level commonsense (like humans vs animals).
 - Intuitive physics properties were useful
 - * Low-level, very general, composable
 - * Higher-level, involves complex mechanisms and tools
 - * Model-based vs model-free reinforcement learning (actions to reward prediction, doesn't make predictions about the environment/world)
 - No explicit prediction about the world, only about the world
 - Adding the simulator gets the higher-level commonsense
 - Expertise-based common knowledge (frame of reference dependent, easier to communicate rather than implicitly learn)

- * Blowing your nose in America vs Japan
- * Explicit vs implicitly learned common sense could differentiate the high vs low level
 - Taught through imitation in animals/humans
- Possibly not the “secret sauce” differentiating humans and animals. Useful for animal-level AGI.
 - * Generalization capabilities and learning could be the “secret sauce”
 - * From common experiences (When an agent experiences it many times, being able to generalize it)
 - * Representations we learn are more generalizable
- Common sense the purpose is for communication
 - * Common ground vs common sense
 - * Maybe tests for common sense should involve interaction with others
- Intelligence tests don’t really guarantee intelligence.
- How do the things discussed in the paper interplay with other modalities?

Journal

- Language is logic with statements?
 - I would say that it makes more sense that logic is done with symbols/representations of concepts
 - Language is more our ability to represent concepts in different modalities (talking/writing) and communicate it (to others)
- Maybe understanding requires consciousness - Does consciousness imply self-awareness? Or is it something “spiritual” that cannot be replicated.
- If understanding cannot be achieved unless the AI is embodied in our world, what happens if we “train” an AGI in our world (embodied in a robot) and transfer the AGI to a computer?
 - It could lose its capacity to reason/understand certain things without the ability to perform actions or perceive in its environment
 - How do we *safely* train an AGI or robot to interact with its surrounding?
 - * It could unintentionally break something or hurt someone, or worse, do it intentionally as a means of exploration
 - * Reinforcement learning and “scolding”?
- You don’t need technical understanding to use things, only functional understanding:
 - Form predictions of interaction results, feel bad if you get the prediction incorrectly, feel good if you get it correct

- * Is this why gambling is fun/addicting?
- “Make instant coffee”
 - Locate coffee jar, locate mug, open coffee jar, pour coffee into mug, check fluent state of mug, locate water, pour water.
 - The order of some of these intentions/actions can be swapped around.
- Utilitarianism is not enough for forming choice preferences:
 - Need some hybrid form of ethics which at least combines deontological, virtue, and utilitarian ethics in some statistical framework.
 - * This system needs to learn new rules as well (deontology)
 - Otherwise we end up with AGI that sacrifice people to save more lives (surgery, trolley problem) or drawing incorrect conclusions (bounded rationality)

2022-02-03 – AGI Readings 10

Computational models of analogy

- Retrieval
 - Integrate both retrieval and mapping?
 - Symbolic-connectionist architecture
 - Different from mapping
 - * Things that tend to look alike tend to have similar causal properties
 - * Mental representations are skewed towards concrete surface properties
- Mapping
 - Base and target representations
 - Structural consistency
 - * 1-1 constraint, parallel correspondences
 - Systematicity
 - * Systems of relations into correspondence are preferred
 - Tiered identity
 - * Identical matches are preferred
 - Matching problem is NP-complete
 - Alignable (electric key vs metal key) vs non-alignable differences (prius vs hummer drivers)

- Abstraction
- Representation
- Generalization
 - Antiunification means finding the least general unifier of 2 expressions
 - * Sort of like the LCF or LCD?
 - probabilityOf(White(Swan)), 0.99
 - probabilityOf(Black(Swan)), 0.01

Abstraction and analogy-making in AI

- Metaphorical concepts
 - Bridge used to cross water, bridge gaps, nose bridge, bridge of a song
 - Many everyday concepts: mirror, shadow, ceiling, driving, sinking
 - AI is too brittle to understand these
 - * Concepts, abstractions, analogy are core things
 - Probabilistic language of thought
- Abstraction and analogy making
 - The “same thing happened to me”
 - Pandemic as another Katrina or war
 - Analogy making is not rare, but a ubiquitous mode of thinking
 - Analogy is key to reasoning, categorization, concept formation, abstraction, and counter-factual inference
- Symbolic methods
 - Structure mapping engine
 - * Mappings made between relations rather than object attributes
 - * Logical propositions and “functions”
 - * SME was able to do well on the raven’s progressive matrices (RPM) problems
 - * SME focuses on syntax rather than semantics
 - Humans are not like this
 - Distinctions between object, attribute, and relation
 - Order of relations, depending heavily on context, people use them flexibly
 - * SME separates representation building and mapping
 - * Also does semi-exhaustive search on matchings
 - Active symbol architecture

- * Copycat concept since analogies are just kinda akin to it
- * Concept network
 - Contains prior knowledge in symbols
 - Symbolic semantic space
- * Workspace
 - Working memory for representations and mappings
- * Perceptual agents
 - Cooperatively and competitively adapt prior knowledge to input situations
- * Temperature
 - Measures quality and coherence of the system's representations and mappings
- Deep learning approaches
 - Word embeddings, high-dimensional vectors, reasoning = numeric operations
 - Too much data required, lack of transparency
 - Dataset matters a lot, can result in overfitting or misleading results
 - DNNs probably aren't actually solving the correct problems
 - Meta learning is more promising
 - * Few-shot learning
 - * Generalization capabilities
 - * Haven't been done for abstraction or analogy
 - * Meta-mapping maps a representation of a task to a related task
- Probabilistic program induction
 - Concept induction
 - * Concepts identified with a program, then perform induction
 - * $L \rightarrow \text{Contains}(L)$
 - Issue is that they need built-in knowledge via program primitives and grammar
- Abstraction and reasoning corpus
 - Few shot learning task benchmark with colored boxes
 - Visual analogies
 - Only innate priors should be the core systems
 - Best submissions were 20% accuracy
 - <https://github.com/fchollet/ARC>
 - <https://www.kaggle.com/icecuber/arc-1st-place-solution/log>
- Discussion

- Symbolic representations and these approaches are brittle
 - * Rely on humans to create/structure prior knowledge
 - * Semi-exhaustive search
 - * Neuro-symbolic approaches
- Active symbol architectures depend on prior knowledge and structure by humans, cannot learn new permanent concepts
- Deep learning doesn't require structured knowledge, but requires large training data and hyperparameter tuning
 - * Often take statistical shortcuts
 - * Meta learning and few shot learning seem promising
- Probabilistic program induction frames concept learning as generating programs, similar to symbolic approaches (also combinatorial search is challenging)
- We should focus on creating AI that can master core systems knowledge
- Evaluate on robustness and accuracy
 - * Evaluate on hidden changing sets of problems (no overfitting)
 - * Evaluate on tasks that require little or no training
 - * Evaluate across multiple domains

Lecture

- Gentner and Forbus
 - Northwestern Psychology/CogSci and CompSci
 - Big prizes in cognitive science/systems
- Analogy - what is it?
 - Comparison between 2 things for the purpose of clarification/explanation.
 - Shows how 2 things are alike with the ultimate goal of making a point about the comparison
 - Analogies that identify identical relationships
 - * Black is to white as on is to off (opposites)
 - Analogies that identify shared abstraction
 - * Human mind as a computer
 - * Solar system as an atom
 - Idioms vs. analogy
 - * Lost the original meaning/analogy for many idioms
 - Bite the bullet
 - Somebody just got bitten by a rattlesnake

- Biting on the bullet lets them persevere through a surgery
 - Bob's your Uncle
 - "And there it is!"
 - Cat got your tongue?
- Analogy involves the comparison of 2 structured representations
 - A familiar base/source domain is mapped to a less familiar target domain
 - Solar system -> atom
 - * Sun maps to nucleus, planets map to electrons
 - * Electrons circle the nucleus
 - Selective mapping
 - * Only some of the things are true
 - Prefer higher-order relations over properties
 - * Causes > color, shape, size
 - Base is a graph/tree
 - * Target is also a graph/tree, but the root node is being inferred
 - Stages in analogy
 - Retrieval
 - * Semantic memory, usually based on surface features
 - Mapping
 - * Determine set of correspondences between elements of base and target
 - * Create a set of candidate inferences that transfer from base to target
 - Abstraction
 - * Leads to an abstraction/schema of similarities of base and target
 - Rerepresentation
 - * Alter representation to improve mapping
 - For example going up or down a hierarchy (dog -> mammal -> live birth)
 - * Encoding of concepts plays an important role in retrieval and mapping
 - * How do the learned analogies affect previously learned concepts
 - SME
 - Stages
 - * Starts with a lot of possible pairs and matches in parallel
 - * Structural consistency is enforced (internally consistent mappings, kernels)

- * Kernels combined into maximal interpretation
 - Might notice something is missing, can make an inference
 - Task independent
 - * Identity + structure
 - * There's no distance knowledge about how close concepts are to each other
 - Embedding spaces of concepts?
 - Structure of concepts encodes semantics
 - * Is-a, has-a, causes, relations
 - SME is only 1-1 mappings
 - * Parallel connectivity
- Possible dimensions of evaluation for analogy mechanisms (***Useful for evaluating other cognitive capabilities in the future***)
 - Match to human data
 - Correctness
 - Computational complexity of retrieval and match
 - Task independence
 - Robustness to concept representations
 - Data-efficient (amount of training)
 - Includes all phases of analogy
 - Integrated with other cognitive capabilities
 - Do any useful inferences come from analogies
- Mitchell paper
 - Quotes from famous people is not evidence nor explanation
 - * "Without concepts there can be no thought, and without analogies there can be no concepts"
 - There's concepts that are bottom-up from experiencing the world
 - What is "easy" for you is not always universal
 - * Bridge metaphors ("Bridge of a song", "Bridge loan")
 - Professor Laird didn't understand this since he doesn't have music theory
 - * ***Have to have rich representations of concepts in order to understand analogies***
 - Are there very different levels or types of processing under analogy?
 - * Abstraction is not just analogy, there's also dropping out things, etc.
 - * We don't make an analogy every single time our current situation reminds us of a past one

- * “Analogies underlie our abilities to flexibly recognize new instances of visual concepts such as a ski race or protest march”
 - DNNs don’t make analogies (a bunch of weights, looks at features, spits out a prediction)
- Is *useful/purposeful/intentional* memory retrieval just analogy?
 - * Depends on the model of memory
 - * **Analogies teach us something new about something from mapping to new concepts**
 - * Memories are just retrievals of the past
 - * Transfer learning / few-shot learning
 - * Memory retrieval is subsumed by analogy
- Discussion
 - Map analogy onto Newell’s time scales (Provide examples!)
 - * Is there a single mechanism for both low-level comparisons and high-level analogies?
 - There’s no one-fits-all analogy mechanism, depends on the modality, context, etc.
 - * Which level for which processing part of analogy (retrieval, mapping, etc.)?
 - Grammars could be considered analogies is automatic
 - New words or grammatical structures is difficult
 - Depends on amount of experience you have
 - * Is retrieval, mapping, etc. automatic or deliberate, or a combination?
 - Depends on the context
 - Some pop up in nature, some you are focusing on (Mapping patterns in nature vs. RPM)
 - Retrieval is probably faster than mapping, especially mapping of new or unseen representations
 - * Is there an analogy module in AGI?
 - Different analogical systems for different things
 - More efficient memory retrieval modules
 - * Is analogy innate or learned or emergent?
 - You have to learn the representations
 - You can learn analogies, first time seeing analogies it’s slow
 - But the system itself is not.

Journal

- What is the difference between analogy and retrieval?

- Storage space is much less of a constraint compared to computing power when translating the capabilities of human brains into computers.
 - Brains can store around 2.5 petabytes of information
 - * A server which costs about \$100,000
 - GPT-3 has around the same number of neurons and synapses as the human brain (80-100 billion neurons)
 - * A cost of around \$10-20 million to train.
 - One approach could be storing highly detailed representations, but ignoring some of these details upon retrieval/analogy. This would be akin to abstracting away the finer details and would speed up computation.
- Analogies can be seen not just as mappings, but also as shortcuts for connecting concepts together.

2022-02-10 – AGI Readings 11

Metacognition for a common model of cognition

- Cognition is the cycle of perceiving, deciding, and acting
 - Many processes such as attention, reasoning, learning, planning, imagination, conscious access, and natural language
- Metacognition is cognition about cognition
 - Reasoning about reasoning, reasoning about learning, learning about reasoning
 - Cognitive process about another structure or cognitive process
- Category 0
 - Perception process -> action process
 - Normal cognition
- Category 1
 - Decision process -> decision process
 - Model-free vs model-based reinforcement learning
 - * Determines amount of influence MB vs MF RL should have. Chooses the more reliable to control human behavior.
 - Self-representation

- * Arises from cognitive processes. Enables some important abilities.
- Reflection and self improvement
 - * Self-reflection, self-modification, improvement
 - * Develop new provisional heuristics to gradually choose the most reliable of them in decision making.
- Self control
 - * Related to health, weight-loss. Lower the value of junk food.
- Category 2
 - Decision process -> action process
 - Do more than control/modulate lower-level behavioral control systems. They are behavioral-control and problem-solving systems.
 - Social cognition
 - * Game theoretic perspective
 - * Consider possible scenarios with our own agent properties and other agents
 - Beliefs, goals, intentions
 - * Social rules (norms, conventions, laws)
 - * Coordinate with other people (organizing meals, working, raising a family)
- Category 3
 - Perception process -> decision process
 - Feedforward representations of environmental stimuli.
 - Context and abstract task relevant information
- Components of metacognition
 - Monitoring
 - * Receives input from cognitive process it attempts to influence
 - Evaluation
 - * Evaluate the received activity/input
 - Planning
 - * Assessment of future success and identifying the best action policies
 - Mental simulation
 - * Play out imagined scenarios before an action is chosen. Rich mental models
 - * Isn't this a subset of planning?
 - Control
 - * Coordinating activities of behavioral-control systems.

- * Arbitrating among systems
 - Multitasking, scheduling, task switching
- * For category-1 processes, it's expected to modulate or bias the behavioral-control system
- Other
 - * Understanding, awareness, generating, organizing, maintaining, modifying, debugging, healing, configuring, adapting
- Consciousness
 - How to explain 1st-person subjective experience of human consciousness?
 - How and why people can experience love, colors, self-doubt
 - Metacognitive observations to achieve consciousness
 - * Observation of external environment
 - * Observation of self in relation to environment
 - * Observation of internal thoughts
 - * Observation of time (past, present, future)
 - * Observation of hypothetical or imaginative thoughts
 - * Observation of observations
 - Affective processing (highly integrated bottom-up signal with top-down conceptual understanding)
- Design issues and case studies
 - Homonculus fallacy
 - * Little person inside brain reasons, central module for control
 - Internal languages of thought
 - * Generate recursive metacognitive processes and outputs
 - * Logical operands, equations, etc.
 - Limitations of human cognition
 - * How well the design mirrors human abilities?
 - Maybe we should not create cognitive systems like this, since there are logical traps that people fall into?
 - I think not, it's better to have a mutual understanding of logical errors than a disconnect of understanding...
- CLARION
 - Uses multiple metacognitive criteria to decide how to use symbolic and subsymbolic processing

- * Whether to use reinforcement, supervised, unsupervised, combination
- * Reasoning mechanism (rule based, similarity based)
- LIDA
 - Based on classification of levels of control

Metacognition in computation: A selected research review

- Statistical model of reasoning
- Multifaceted theory of mind
- Monitoring
 - Judgements of learning
 - Feelings of knowing
 - Confidence in answers
 - Knowledge acquisition, retention, retrieval
- Sources of activation confusion
 - Intersection of 2 or more semantic nodes triggered by the terms of a given question
- Self-explanation effect
 - Strong and positive effect
 - Good old fashioned AI is explicit representation
- Human understanding is the process of executing some model of the world
 - Not sure about this one, I think there's more to it, such as the reasoning, and learning of processes. Also self-introspection.
 - Logical system can answer queries about world
 - **Humans learn quickly by ascribing beliefs and goals to a machine than by analyzing and explaining it in terms of code and states**
- Deduction structure
 - Mathematical abstraction of belief systems
 - How to generate beliefs?
 - * Reason forwards to figure out what action/computation to perform
 - * Reason backwards to figure out what what metacognitive monitoring/reasoning to explain a failure or learn
 - Bounded rationality

- * Utility functions
- Monitoring computations also takes time and cost
 - Motivates a real-time scheduler?

Lecture

- Computer architectures have lots of levels of abstraction
 - Application -> middleware -> virtual machine -> hardware
 - Each level implements the level above it, uniform and independent of the lower levels
 - This is not metacognition
 - Protect programmer from the lower level details
- Cognitive architectures
 - Architecture layer: implemented in C, Lisp, Java
 - * Implements Soar, ACT-R
 - Architecture is fixed and defines representations and primitive operations on cognitive layers
 - * Should AGI architecture be fixed? Seems like human brain architectures can change
 - Interprets/executes knowledge to produce behavior
 - * Architecture processing is task-independent (no task-dependent tests in code)
 - * Interpretation of cognitive knowledge makes it task-dependent
 - * Implementation language and exact algorithms are irrelevant (except for speed)
 - Cognitive layer has knowledge encoded as rules, semantic networks, etc.
 - Need to discuss metacognitive layer
 - * Data structures
 - Code and data structures representing agent knowledge
 - Metadata of retrieval and use of agent knowledge
 - ACT-R is utility
- ***Need a simple model of base cognitive processing to discuss meta cognitive processing***
 - Decide -> Action cycle
 - **External environment** provides **perceptual data** -> **short-term task state**
 - * Some abstractions/configurations as well
 - * Not about agent's processing or problem solving
 - Task state is tested as **procedural knowledge**

- * Put information both into decide and action
- **Metadata: state and procedural knowledge**
 - * Also comes into play into decide
 - “Intrinsic reward is used for learning, not decision making”
 - Actions alter the task state
 - [[cognitive-basics.JPG]]
 - So there’s perception, task state, metadata, procedural knowledge, deciding, acting, monitoring, learning, modifying parameters
- Decide
 - Decision could be random, independent of state and goal
 - Mapping from state, metadata, and goal to action
 - Evaluates proposed actions in context of state and goal
 - NN, decision trees, rules, etc.
 - Meta stuff
 - * Evaluate proposed actions using model of their effects
 - Look-ahead search
 - * Use historical information about similar applications
 - Retrospection
 - * Choosing to allocate resources to things
 - **Metacognition converts to normal cognition over time**
 - * System 1 vs. 2
 - * Why things get faster over time
 - Prioritize exponential learning over immediate performance
 - * Postulate that in new environments/tasks, more resources are allocated to metacognition in order to better generalize things learned from other areas
- Issues in metareasoning
 - What are advantages/disadvantages?
 - When is it invoked?
 - What data is available for metareasoning?
 - What additional data and processing are needed to support metacognition beyond the simple design/loop provided?
 - How do results of metareasoning influence base level processing?
- Decide metareasoning examples
 - Internal planning using models of own actions

- Counterplanning against opponent using model of their processing to predict their actions
- Retrospective analysis of success/failure of prior decisions
- Advantages
 - * Better quality decisions
 - * Better utilization of time
 - * General learning
- Disadvantages
 - * Less reactivity of environmental dynamics
 - * Metareasoning might not have access to all metadata used in fixed decision
 - Can be costly to expose metadata to metareasoning
 - * If knowledge of self is inaccurate, could be worse than a simple decision
 - “Overthinking” vs “intuition”
 - * Need to avoid homunculus fallacy
 - Can’t be meta all the way up
 - Infinite recursion of metareasoning
 - * More complex architecture
 - Computational and memory overhead
 - * Requires additional metaknowledge for metareasoning
 - AGI could bug out
- ***Should AGI have perfect knowledge?***
 - It would be very computationally and memory expensive
 - Do people forget things based on space constraints? Or from interference during retrieval?
 - * Neural network will only hold so much information, will begin to degrade after exceeding certain capacity.
 - * Forgetting may be better to generalize to new situations
 - Not really forgetting, but merging and abstracting
- When is metacognition invoked?
 - Always used
 - * Only practical if environmental dynamics take longer than internal processing. Such as game playing programs
 - * We have downtime during chill tasks and sleeping (plenty of time to do metacognition)
 - When metadata suggests it will be useful
 - * Insufficient metadata

- Maybe can be thought of as sample complexity and such
- Might do metareasoning to generate more data?
- * High variance in metadata
- * Not enough experiences
- * Surprise emotion
- * Nothing else to do at the moment
- Task-specific knowledge monitoring decision process
 - * Requires decision process to be incorporated and not a black box
- Explicit decision
 - * Task action to decide to go meta
- Don't need to use metacognition once we've learned how to walk or run for example
- Metacognition gets compiled down to cognition
- Data available for metareasoning
 - State of task
 - Metadata for action selection (state of decision, options, evaluations, rewards)
 - Internal model of own actions and environment
 - * Supports planning and internal imagining
 - * Have a model of "self"
 - This is problematic and expensive
 - * Instead, can execute simulation of "self" in hypothetical situations
 - "What would I do if xyz?"
 - Historical information of prior problem solving
 - Additional data and processing
 - * Represent metadata explicitly or declaratively?
 - * Represent hypothetical task state and modify without interference with base-level representation or reasoning
 - For planning, imagination, etc.
 - **Need to develop an imagination space separate from actual space**
 - Still needs to be somewhat hooked up to actual systems
 - * Metareasoning procedural knowledge that is task independent
 - * Means to invoke base-level processing, apply knowledge to world and actions
 - Need to also be able to apply it in non-actual spaces
 - * Expectations and surprise
 - * Historical information

- Module or recursion?
 - Separate module
 - * Clarity in meta vs. base reasoning
 - * Separation avoids risks of interference
 - * Architecture for metareasoning can be customized
 - * Easier solution to homunculus fallacy
 - Recursive processing
 - * Fewer additional mechanisms
 - * Need a way to avoid interference
 - Separation of meta and real life
 - * All reasoning is the same, so it's easier to intermix base and meta
 - Using natural language, analogy, memory access
- How do results influence real-life processing?
 - Provide input to decision, use best choice
 - Detect error in internal models/knowledge and correct them
 - Create/tune knowledge that kicks in to influence future decisions
 - * Realize during meta-planning that you need to correct a future decision
 - * Current model of self suggests you will make a mistake
 - * “Pick up the milk on the way home from work”
- Accessing internal long-term memory
 - Reason about intermediate results from long-term memory
 - * Feeling of knowing, tip of tongue, ease of learning
 - It kinda feels frustrating, but after getting the result, it feels good
- Rant for the day
 - Meta-learning is not pre-learning
 - * Few-shot learning, transfer learning
 - Meta-learning is learning about how to make learning better
 - * Learning strategies

Discussion

- What are other examples of metacognition?
- How to apply to neural network models?

Journal

- Represent hypothetical task state and modify without interference with base-level representation or reasoning
 - For planning, imagination, etc.
 - Need to develop an imagination space separate from actual space
 - * Like a digital twin of the real world
 - * Knowledge and learned interactions still need to be somewhat hooked up to actual systems (transfer learning/knowledge)
- What are other examples of metacognition?
 - Conscious scheduling of tasks
 - Management of goals
- How to apply to neural network models?
 - Meta-learning, hyperparameter tuning, automatic weight generation, neural architecture search
- Monitoring computations also takes time and cost
 - Motivates a real-time scheduler?
- There seems to be a similar overall structure in many of these meta-level systems (metacognition, emotion appraisal theory, predictive inference, learning systems). Generally it seems to be: perceive environment -> store knowledge -> evaluate internal state -> make decision -> perform action -> reevaluate environment and internal state.
 - Maybe we can leverage this by creating only 1 overarching system that does this perception action loop.
 - Then we can wrap the other meta-level systems around this

2022-02-11 – AGI Readings 12

On the functional contributions of emotion mechanisms to (artificial) cognition and intelligence

- Emotion and cognition are closely intertwined
- Giving emotions to machines could be different from giving machines the ability to recognize emotions

- 3 positions
 - Emotions have no functions
 - Emotions once served functions that are no longer necessarily appropriate
 - Emotions serve important functions now
- Intrapersonal
 - Homeostatic regulation
 - * Hormonal changes for emergencies (adrenaline)
 - Cognitive override
 - * Emotions can sometimes jumble our rational thoughts
 - * They can orient attention and biases for action selection and path search
 - * Basic emotions (sadness, happiness, disgust, anxiety, anger)
 - Manifest as cognitive biases towards things
 - Behavioral adaptation
 - * Reward/punishments for immediate consequences and anticipated consequences
- Interpersonal
 - Communication
 - * Orthogonal emotional states are expressed very differently
 - Social exchange
 - * Facilitates social harmony
 - * “Guilt” - social engagement aimed at reconciliation
- Emotional components for AI
 - Role of embodiment
 - * Essential control variables needed to operate homeostatically
 - Blood glucose levels, battery levels, etc.
 - Expressing and recognizing emotions
 - * AI should be able to interpret and recognize human emotional expression as well
 - * How believable is the emotional expression by a robot/AI?
 - Feeling emotions and higher-level cognitive functions
 - * The feeling of emotion is a physiological change and sensations accompanying them.
 - * Sometimes, we lose control over emotions and thought processes.

Emotion in the common model of cognition

- Emotions emerge at the biological level, act on cognitive and rational control, and are expressed/interpreted at the social level
- Theories
 - Appraisal theory (personal significance attributed to situational factors)
 - Core affect theory (emotions explained in simple terms without needing emotion labels)
 - Somatic markers (emotional tags attached to information)
 - Primary-process affect (lower level processes combine with secondary-level processes (e.g. memory))
- Functions
 - Alarm and interruption
 - * Interrupts for ongoing processes
 - * Changes in attentional control
 - Results in redirection or acceleration
 - Procedural reward
 - * Valuation of rewards and punishment are emotional
 - Social emotionality
 - * Competition, collaboration, assistance
 - * Moral schemas
 - Relations of trust, subordination, competition, etc.
 - Neurophysiological plausibility
 - * Simulates effect of homeostasis
 - Changes in affect and behavior
- Emotions in the common model
 - Structure and processing
 - * Powerful heuristic in bounded rationality
 - Surprise and desirability for bottom-up and top-down inputs to attention (abstraction during memory retrieval)
 - * Symbolic information
 - Memory and content
 - * Somatic markers map onto this
 - * Symbolic structures (memory chunks) and sub-symbolic quantities

- Learning
 - * Emotional structures and metadata should be learnable
 - * Model should eventually learn emotional values associated with different memories

Emotion and Decision Making

- Potent drivers of decision making
 - Bounded rationality is incomplete, emotion exists
- Integral emotions influence decision making
 - Anxiety about risky choices causes the choice of the safer option
 - Gratitude and donating
 - Is reason a slave to emotion, or is emotion a slave to reason?
 - Financial investments and euphoria
 - How do we model this? As bias? Or as a guide?
- Incidental emotions influence decision making
 - Can carry over from one situation to the other
 - Could be used as a strong bias for decisions
 - Moderating/coping factors
- Emotional valence is only 1 of several dimensions that shape emotions' influence on decision making
 - Valence/sentiment is just 1 of many dimensions
 - Appraisal tendency framework
 - Emotion could trigger cognitive predisposition to assess future events in line with the appraisal dimensions that triggered the emotions
- Emotions shape decisions via the content of thought
 - Anger scores high on the dimensions of certainty, control, and others' responsibility and low on pleasantness
 - What types of thinking and reasoning we'll do
- Emotions shape decisions via the depth of thought
 - How much reasoning and thinking we'll do
 - Excessive rumination is not good
- Emotions shape decisions via goal activation

- Emotions serve an adaptive coordination role
- Sets of responses that address problems quickly
- Emotions influence interpersonal decision making
 - Optimal navigation of social decisions
 - Effects of emotions are qualified by contextual variables
 - Game-theoretic games played in multiple iterations allows for emotion to communicate and correct things
- Unwanted effects of emotion on decision making can be reduced under certain circumstances
 - Let time pass before making a decision (emotions are typically short lived)
 - Coping/suppression mechanisms
 - Reappraisal of the situations
 - Increase cognitive effort through financial incentives, crowding out emotion, increasing awareness of misattribution, choice architecture
 - * Cognitive self-awareness
- General model
 - Emotion-imbued choice model
 - Uses current emotions, incidental influences, characteristics of decision maker, characteristics of options, conscious and/or nonconscious evaluation, decision, and expected outcomes

Lecture

- Emotion theories
 - Circumplex models
 - * Intensity and valence
 - Like sentiment and magnitude
 - * 2 dimensions don't seem to be enough
 - Appraisal theory
 - * Emotion depends on current appraisals of emotion
 - Usually 2 dimensions, sometimes up to 14 dimensions
 - Can have continuous values
 - * Mood is longer term, decaying emotions
 - * Feelings are the agent's perception of emotion and mood that are exposed

- Includes interpretation of emotion (fear, joy, disappointment)
- Appraisal theories
 - Situation + goals
 - Appraise the state depending on:
 - * Novelty, goal relevance, goal conduciveness, expectedness, causal agency, etc.
 - * Anger = loss of causal agency and in a conducive situation
 - Produce emotion based on appraisals
 - Cope with the emotion
 - * Do things to change the situation/goals or your perspective on it
 - * Uses appraisals to figure out how to return to the baseline
 - * Do bipolar people struggle with this coping mechanism? Do the emotions produced by their appraisals become very extreme?
 - * Meta coping
 - Problem focused: act in the world to change situation
 - Emotion focused internal action (change goals, beliefs, redescribe) that leads to suppression
 - “I’m sorry you didn’t get into graduate school at Stanford”
 - “Oh I didn’t even want to be in California and also they work students too hard.”
 - Have a set of N innate appraisals. We learn how to interpret the situation to fit these innate appraisals.
 - * Example appraisal dimensions
 - Suddenness
 - Abruptness of stimulus. Does not depend on what you know.
 - Unpredictability
 - Could not be predicted. Depends on what you know.
 - Intrinsic pleasantness
 - Extent to which stimulus is pleasant (independent of goal)
 - Relevance
 - Importance of stimulus relative to goal
 - Causal agent
 - Who caused the stimulus
 - Causal motive
 - Motivation of the causal agent (perceived)
 - Outcome probability
 - Probability of stimulus occurring
 - Discrepancy from expectation

- Stimulus did not match prediction
- Conduciveness
- Stimulus is good/bad for the goal
- Control
- Anyone can influence the stimulus
- Power
- Agent can influence the stimulus
- * [[emotion-dimensions.JPG]]
 - Hypothesis is that the input based on what your goals/knowledge are can be tuned to different cultures/communities
- Simple model (Reinforcement learning, use feelings as intrinsic reward)
 - Perception
 - * Get perceptual state
 - Low level appraisal
 - * Other processing in parallel
 - * Derives internal state, situation & goals, feelings and appraisals
 - * Can do a prediction search for situational assessment of hypothetical states/goals/appraisals
 - Can be used to determine possible actions
 - High level appraisal
 - * Causal agent/motive, discrepancy, conduciveness, control/power
 - Can do actions
 - * Possible actions suggested by procedural knowledge, internal search, etc.
 - * Have action tendencies based on these appraisals
 - * Attempt to cope
 - [[rl-emotion.JPG]]
- Impacts of emotion
 - Systemic preparation
 - * Physiological changes
 - Influence decision making
 - * Dimensionality reduction of state to a smaller set of relevant dimensions
 - * Suggestions of actions
 - * Evaluation of alternative states
 - * Attempts at coping

- Learning
 - * Component of reward in RL
 - * Associated with semantic memories and episodic memories
- Communication
 - * Extra channel of information about agent state and goal progress
 - * Help navigate and coordinate social interaction by providing information about others' beliefs, motives, intentions, and dispositions
- Why have emotions in AGI?
 - Faster processing of information/state. Heuristics or emotions?
 - * Do we need the final mapping to joy, fear, anger, etc.?
 - Robots still need to be able to communicate emotion to humans
 - Theory of mind mapped onto humans
 - Most human interactions have some emotion
 - Infants need to express emotions so that caretakers will tend to them

Journal

- Can computers “feel emotion”?
 - Our brains send signals to distribute chemicals depending on our intuition and appraisal of a situation. These chemicals make the brain more receptive to pleasure (joy) or make the perception system in the brain more alert (fear).
 - Computers can likely emulate this experience through appraisal and reward systems. Computers may also have sensors that can interpret power usage, heat, fan speed, clock speed, etc. This could be seen as analogous to human's heart rate, sweating, etc.
- Why have emotions in AGI?
 - Grounding in human-like emotional experiences.
 - Ability to efficiently interpret and convey emotions to/from humans.
 - Cultivate trust in AGI - human interactions (don't want to create AGI sociopaths)
 - * Ability to empathize with humans and other AGI (crucial for social interactions and ethical reasoning)
- Emotions have to function in real-time
 - In order to properly influence reward mechanisms and reinforcement learning.
 - This means there should be a fast/intuitive appraisal mechanism and a slower reasoning appraisal mechanism.

- Beyond just appraisal theory
 - It helps people to further abstract/simplify emotions into either positive or negative (and magnitude).

2022-02-16 – AGI Readings 13

Learning fast and slow: Levels of learning in GAIA

- Learning can be split into 2 levels
- Level 1: Architectural Learning Algorithms
 - Learn no matter what
 - We cannot explicitly invoke or inhibit these learning mechanisms
 - Operant conditioning, classical conditioning, habituation, sensitization, rote learning
 - No restrictions on types of knowledge representations the L1 algorithms learn over
 - * Can learn over perceptual data, high dimensional representations, symbolic representations, etc.
- Level 2: Deliberate learning strategies
 - Learn only if there's time to deliberately improve
 - ***These are when learning becomes the goal/motivation of the agent***
 - * Learning strategies, not information gathering strategies
 - Unique to humans?
 - * Monkeys learning tasks (don't they sometimes decide to just practice tasks unprompted?)
 - * Dogs/dolphins learning tricks
 - Examples
 - * Students using flashcards to train and test themselves
 - * Athlete practicing 3-point shots
 - Cognitive capabilities
 - * Analogy, attention, decision making, dialog, goal-based reasoning, spatial reasoning, temporal reasoning, theory of mind
 - Strategies:
 - * Spaced practice (spacing out studying)
 - * Retrieval practice (memory retrieval without aids)
 - * Elaboration (explaining ideas with many details)

- * Interleaving (switching between ideas while studying)
- * Concrete examples (using specific examples to understand abstract ideas)
- * Dual coding (combining words and visuals)
- Identifying learning opportunities
 - * Missing, conflicting, or uncertain knowledge
 - * Unexpected event or novel situation
 - * External knowledge available to learn from
 - * Past experiences to enhance future performance
- Level 0: Population evolution
 - L2 mechanisms are learned through L1 processes
- Level 1+: Innate strategies
 - In natural systems, there's innate drives that direct behavior
 - Tendency for youth to play, imitation
- Level 2+: Social strategies
 - Agent can learn strategies and abilities from other agents
- Level 3: Deliberate strategies from modifying L1
 - Behaviors that try to change the physiology of learning mechanisms in the brain
 - Resting, exercising, drugs, meditation, etc.

The computational gauntlet of human-like learning

- Machine learning has become very uniform. AI community should devote more energy to other paradigms that hold more potential. Human-like learning, for example is different than machine induction
- Children have a curriculum which takes years, but they are still able to generalize easily. For example, autonomous vehicle driving takes millions of miles.
- Symbolic structures (condition-action rules, decision trees, grammars,)
- Neural networks are very far from human learning.
 - How to fix the sample size learning issue?
 - Sometimes they can encode certain spatial relations and abstract relations (graph convolutional networks)
 - K-shot learning
- ***Learning involves acquisition of modular cognitive structures***

- Dynamic cognitive architectures that are self-managed
- Cognitive structures are relational. They refer to connections between entities and events
 - * Representational systems that merge semantic and episodic memory
- Cognitive structures are acquired and refined rapidly from small numbers of training cases
- **Learned cognitive structures can be composed during performance**
 - Online learning system
 - Learning is an incremental activity that processes 1 activity at a time
 - Expertise is applied piecemeal (1 element at a time)
 - * However, the lower level systems are incredibly parallelized. So, it depends on the cognition band level.
- **Learning is guided by knowledge that aids interpretation of new experiences**
 - Using knowledge to improve learning and vice versa
- Statistical learning should involve both the generation and testing of hypotheses.

Lecture

- Summary on emotion:
 - Useful for decision making, learning, and interaction
 - Suggests a general value function useful in planning and learning
 - Big question is how to compute the task-dependent aspects of appraisals
 - Some appraisals
 - * Easy to create task independently
 - Suddenness, unpredictability, intrinsic pleasantness, outcome probability, discrepancy from expectation, conduciveness
 - * Difficult to compute task independently (depends on task)
 - Relevance, causal agent, motivation, control, power
- Authors
 - Pat Langley
 - * Psychology, machine learning, cognitive systems
 - Shiwali Mohan
 - * PhD UMich
 - * Xerox PARC

- * Human-aware AI
- Skill learning in humans
 - Acquisition
 - * Declarative
 - Consolidation
 - * Declarative and procedural
 - Tuning
 - * Procedural
 - [[skill-learning.JPG]]
- Langley paper: Human learning vs. Current ML/DL
 - Learning is online and incremental
 - * Expertise is learned one element at a time
 - Acquires modular relational structures
 - * Learning can be guided by prior knowledge
 - * Cognitive structures can be composed during performance
 - Contrast to batch learning
- Newell's time scale of human action
 - Emphasis on statistical learning is neural circuitry and neurons band
 - Learning new symbolic/relational structures is on deliberate act band
- Questions for AGI
 - What is missed if you only learn at the symbolic level?
 - * Miss ERM learning over data
 - What is missed if you only learn over existing structures?
 - How to learn new structures quickly from learning at the neuron and circuitry level?
 - * How to get immediate (quick) learning of new concepts?
 - * How to get influence of previously learned concepts?
 - * How to get composition of existing concepts?
 - Do hybrid approaches provide solutions?
 - Can AGIs take other approaches to learning?
 - * Not always incremental, also having offline batch learning?
- 2 kinds of learning

- Architectural learning mechanisms
 - * Incrementally captures knowledge from ongoing experiences
 - * Classification, categorization, sequences, reinforcement
- Knowledge based learning strategies
 - * Create experiences for L1 mechanisms to learn
 - * Arbitrary processing using arbitrary task knowledge at limited time
 - * Practice and repetition, deliberate study, experimentation, playing
- L1
 - Captures knowledge from an agent's experiences
 - * Perception, internal data, and metadata
 - Automatic
 - * Innate, effortless, online, always active
 - * Does not compete with task reasoning
 - Diverse learning mechanisms for long-term memory stores
 - * Perceptual, semantic, procedural, episodic
 - * Instances, categories, metadata, procedures, sequences
 - Limitations
 - * No deliberate learning
 - Can't decide to learn something
 - Suggests need for meta-cognitive control
 - * Constrained to incremental constant-time learning algorithms
 - Suggests need for unconstrained learning analysis
 - * Learns from data recently accessed by task processing
- L2
 - Deliberate
 - * Initiated and controlled by agent knowledge
 - * Tasks that compete with other task reasoning
 - * Can occur during slack time between tasks
 - Control external experience
 - Enrich experience with additional knowledge & reasoning
 - * Interaction with other agents
 - * Recall, replay, analyze prior experiences
 - * Retrieval and analogy

- * Planning, hypothesizing
 - Strategies can be learned and improved with experience
 - If you don't learn certain tasks deliberately enough (math, coding, etc.), then you become too high level and take too long.
- L0
 - Population evolution, creates L1 mechanisms
- L1
 - Architectural learning mechanisms
- L2 (***The future of AGI***)
 - Innate learning strategies
 - * Curiosity, imitation, play
 - Knowledge-based learning strategies
 - * Probably not in animals
- L2+
 - Social learning strategies
 - * Organized education, funded research, conferences
 - Probably not in animals
- L3
 - Modifications of L1 strategies
 - * Rest, exercise, ingesting drugs, meditation

Discussion

- What are other important level 2 strategies?
 - Ask somebody for an answer
- How to get level 2 strategies into AGI systems?

Journal

- "Expertise is applied piecemeal (1 element at a time)"

- However, the lower level systems are incredibly parallelized. So, the level of focus/parallelization really depends on the cognition band level.
- “The dangers of sampling bias are well known”
 - This is a feature not just unique to neural networks or statistical learning, but also to humans!
- “Airplanes do not fly like birds, so why should computers think or learn like people?”
 - I agree with this. Why do we need to “start from scratch” when creating an AGI? Can we not leverage the advances in statistical learning to jumpstart the process? For example, one can use a pretrained DNN to handle perception-based tasks such as scene segmentation, object detection, classification, NLP, etc. It is unreasonable to restrict ourselves to creating an AGI that needs 5+ years to learn.
 - Computers are also fundamentally different from people, so we should leverage advantages and disadvantages that computers have (lots of memory and storage, fast operations, not great at parallelization)
- Is it unreasonable to try and incorporate deep learning and machine learning as low-level *components* of a human-level learning/AGI system?
 - These would be neural circuitry or cognitive level
- How to learn fast?
 - Use pretrained networks and few-shot learning.
 - Transfer learning from one domain to another.
 - Symbol systems, operations, and rules are fast.
- How to leverage knowledge in new learning?
 - Imagination and generation of new scenarios (like problem search but for learning).
 - Apply concepts learned from one domain to meta learning (e.g., see people who ask questions get praised -> ask more questions).
 - * Another example: learn about real-time systems -> use real-time scheduling strategies when studying -> learn faster
 - * These are more level 2 learning
 - Recognize and retrieve previous examples that have some similarity to new instance being learned

Links

- [[]]