

Object Relationship Spaces for AI-ML: A Framework for  
Clearly Defined, STEM-Compatible, Project-Level, Functional Units of "Intelligence"  
For AI Design, Analysis, Performance, Architecture, and Operating Systems

*Geoffrey Gordon Ashbrook, 2023.03-04*

**Abstract**

*There is a need for the use of well defined performance frameworks to describe the goals and skills/abilities of systems including AI.*

The overall agenda here is to move toward clearer communication and better definitions, including the pragmatic utilization of universal intersecting/interlocking areas.

*This proposed object-relationship-space framework can be used for guiding project-specific system design, goal-setting, discussion, testing, analysis, reporting, regulation, documentation, etc.*

*For more detail on what is meant by 'design': to manage and enable smaller or larger scale AI projects coordinating required abilities across internal and external components, including "symbolic" logistics and "sub-symbolic" training (including for AI-self-management), and whole operating-systems for AI.*

AI must be able to handle "objects" in the following interlocking contexts:

1. object-relationship-spaces
  2. (internal/external) project-object-database (in a project-framework)
  3. project-participants (in a project-framework & participation-space
- such that 'objects' are defined as existing outside of the AI in an overall project context, and that so long as the AI effectively deals with these project-objects across these contexts including internal and external handling, it does not matter how the AI 'internally' handles objects. For example, alternative methods of internal handling/processing/management include:
- symbolic vs. sub-symbolic
  - single or end-to-end vs. multiple or ensemble or hybrid
  - parametric, nonparametric
  - explainable vs. black-box
  - higher dimensional vs. lower dimensional
  - calculation vs. intuitive pattern recognition
  - similar-to-h.sapiens, vs. not similar to h.sapiens
- etc.

*This being a subset of a study of System & Definition behavior Studies, two useful sets of connections include **Interconnecting/Intersecting Areas** and **Definitions, Frameworks, and Participation**.*

***Interconnecting/Intersecting Areas:***

*A repeating theme, context, and agenda in this paper is to pragmatically leverage the interconnected functionality of clear definitions, STEM, projects, participation, coordination, system-fitness, positive values, and productivity.*

*To reiterate and state this as clearly and openly as possible, the context and agenda here is a project, best practice, positive-values, productivity, context.*

***Definitions, Frameworks, and Participation***

*1. Discussion with undefined terms (for example specific abilities) can loop indefinitely regardless of the abilities of AI (at that time or in the case of changes over time in what AI can do). Undefined & under-defined goals, terms, and definitions tell us too little about what is needed, what the system can do, and if the system can do what is needed.*

*2. Telser Rule Loops: Where "AI" is undefined and every new development is dismissed as "not real AI," the failure to define "AI" tells us too little about what is needed, what the system can do, and if the system can do what is needed.*

*3. According to an Object-Relationship-Space framework in a project and participation context, AI-ML technology can as of March, 2023 join h.sapiens-humans as a participant in projects, with specific skills/abilities to handle specific project-objects, where projects, participants, and objects in object relations spaces, are clearly and functionally defined in a STEM context. This Object-Relationship-Space framework should define what is needed, what the system can do, and that the system can do what is needed. The details of how the practical context of 'project participation' partially overlaps with the vague context of 'intelligence' are likely significant.*

*Part one concerns a brief overview of the framework.*

*Part two concerns using the framework, e.g. so you can construct your own well defined goals and tests for abilities of AI systems.*

*Part three concerns a discussion of the discussion of AI, e.g. so you can critique statements in what you read about AI.*

*Part four concerns goals and agendas, background concepts and principles, and future design factors.*

*Appendices include more examples and details.*

## **Table of Contents: Brief**

Abstract

A Narrative Introduction in Two Parts

### **Part 1: Framework in a Nutshell**

The Object Relationship Space Framework

- 1.1 Example Object Relationship Space List
- 1.2 Many lists in One
- 1.3 Networked-AI Components

### **Part 2: Using The Framework**

- 2.1 Examples: AI Skills, Comparing Three Chatbots
- 2.2 Examples: AI Skills Mapped to Object Spaces
- 2.3 Adding Levels, Adding Steps, Adding Objects
- 2.4 Hybrid/cross-model skills for AI
- 2.5 Heuristic & Pseudocode for AI Management with Object Spaces
- 2.6 Modularity, Scale, and AI-Component Networks
- 2.7 Model 'Explainability' as 'Explainability, Reliability, and Security'
- 2.8 Mapping a general problems-space for AI & Mind
- 2.9 Object Relationship Based Testing

### **Part 3: Discussing the Discussion of AI**

- 3.1 Definitions of Terms
- 3.2 Data Sources Discussing AI
- 3.3 Examining Tests for AI
- 3.4 What do we do with ChatGPT?

### **Part 4: Goals, Background & Future**

- 4.1 Agenda & Goals
- 4.2 Background Concepts and Principles
- 4.3 Future Design Factors

## **Appendices**

## Table of Contents: Detailed

Abstract

A Narrative Introduction in Two Parts:

Introduction Part 1. Chess in Blade Runner

Introduction Part 2. Defining AI Goals and 'Objects'

### Part 1: Framework in a Nutshell -> The Object Relationship Space Framework

1.1 Example General Object Relationship Space List

1.2 Many lists in One

1.3 Networked-AI Components

### Part 2: Using The Framework

2.1 Examples: AI Skills, Comparing Three Chatbots

2.2 Examples: AI Skills Mapped to Object Spaces

2.3 Adding Levels, Adding Steps, Adding Objects

2.4 Hybrid/cross-model skills for AI

2.5 Heuristic & Pseudocode for AI Management with Object Spaces

2.6 Modularity, Scale, and AI-Component Networks: and the need for a literal or proverbial operating system.

2.7 Model 'Explainability' as 'Explainability, Reliability, and Security'

2.8 General Mind-Space: Mapping a general problem-space for AI

2.9 Object Relationship Based Testing

### Part 3: Discussing the Discussion of AI

(In Summary)

3.1 Definitions of Terms

3.2 Data Sources Discussing AI

3.3 Examining Tests for AI

3.4 What do we do with ChatGPT?

(In Detail)

3.1 Definitions of Terms

3.1.1 Terminology Issues 1: The tangled Semantics of human ability.

3.1.2 Terminology Issues 2: What has been defined or is not-defined?

3.1.3 Terminology Issues 3: Navigate Jargon Pragmatically

3.1.4 Terminology Issues 4:  
Fictional frames of reference are bad:

3.1.5 Terminology Issues 5:  
Problematic multiple meanings of unavoidable terms

3.1.6 Terminology Issues 6: Beware Non-sequitur Conclusions

- 3.1.7 Discussions of Model "explanation"
  - Double standard between symbolic and subsymbolic
- 3.1.8 Reification
- 3.1.9 Definition Collapse: Maintain your definitions
- 3.1.10 Discussions of Model "explanation"
  - Double standard between symbolic and subsymbolic
- 3.1.10 Potemkin Villages and Telepathy-Tests
- 3.1.11 Terminology and Interpretation of Intent:
  - Azimov's Laws of Robotics
  - ELIZA the Psychotherapy AI
- 3.1.12 Terms that people cannot define while pretending they can
- 3.1.13 local context specific definitions
- 3.1.14 Negative Definitions
- 3.1.15 Indirect Definitions
  - 'fail to disprove the nul hypothesis'
  - value function & meaning
- 3.1.16 Participation
- 3.1.17 Generalization
- 3.1.18 Controversial Topics
- 3.2 What To Read:
  - 3.2.1 The Three-Legged Writing Stool
  - 3.2.2 History
  - 3.2.3 Interdisciplinary Area Recommendations
  - 3.2.4 Do AI Projects
  - 3.2.5 Book Recommendations
  - 3.2.6 Read classic Science Fiction: Back to Blade Runner again
- 3.3 Examining Tests for AI:
  - (Under Construction)
  - 3.3.1 - looking at winograd schemas
  - 3.3.2 - Sally Anne Tasks
- 3.4 Influences on Model Architecture
- 3.5 What do we do with Large Language Models & ChatGPT?

## **Part 4: Goals, Background & Future:**

(In Summary)

- 4.1 Agenda & Goals
- 4.2 Background Concepts and Principles
- 4.3 Future Design Factors

(In Detail)

### **Introduction to Part 4**

#### **4.1 Agenda & Goals**

- 4.1.1 The goals and agenda here
- 4.1.2 Defining your goals

#### **4.2 Background Concepts and Principles**

- 4.2.1 "Intersecting/Interlocking Areas"
- 4.2.2 input output measures...or next section
- 4.2.3 Higher Dimensional Frontier: Tensors & Matrices

- 4.2.4 Projects & Project Context
- 4.2.5 Instrumentalism and Realism
- 4.2.6 Big Other Areas
- 4.2.7 Gamification
- 4.2.8 Ambiguous Equivalence
- 4.2.9 Errors and Mistakes

### **4.3 Future Design Factors**

- 4.3.1 Biology
  - 4.3.1.1 Integration with biological systems
  - 4.3.1.2 Use or imitation (of biological functions)
  - 4.3.1.3 Compare and contrast for study and understanding
  - 4.3.1.4 Highlighting known areas of development
  - 4.3.1.5 Highlighting still not well understood areas
  - 4.3.1.6 Highlighting predictable problems
  - 4.3.1.7 The science of sleep
  - 4.3.1.8 The science of memory
  - 4.3.1.9 Non-chordata "intelligence" & decision making
  - 4.3.1.10 Science of Mind/Cognition/Consciousness
  - 4.3.1.11 Science of Entheogens
  - 4.3.1.12 Science of Mindfulness
  - 4.3.1.13 Bio-Nano-Coded AI
  - 4.3.1.14 AI in synthetic organisms for terraforming
  - 4.3.1.15 DNA/RNA based digital information interface
  - 4.3.1.16 The Dragon Project (hybrid synthetic chimeras)
- 4.3.2 Cybersecurity and AI
- 4.3.3 Quantum information theory & under-the-hood optimizations
- 4.3.4 "Generalization" vs. Deployment
- 4.3.5 Nanotech
- 4.3.6 Understanding Exponential Elbows
- 4.3.7 'Complexity' Nonlinearity Dynamical and Systems Sciences
- 4.3.8 Ethics, Projects, Best Practice & STEM
  - STEM, Ethics & Mindfulness
- 4.3.9 Projects: Agile
- 4.3.10 AI and Code Testing
- 4.3.11 The Long Term Memory Storage Problem
- 4.3.12 The challenge of orientation and navigation in mind-space
- 4.3.13 Human machine interactions, biology machine integration
- 4.3.14 Project-Context Decision-Making Involving Participants
  - and Components
- 4.3.15 Question Space
- 4.3.16 Self-Awareness Apace
- 4.3.17 Analogies
- 4.3.18 System Epidemiology
- 4.3.19 The Cambrian Midway Point
- 4.3.20 Parent-Child Policy Decision
- 4.3.21 Culture & AI
- 4.3.22 Kasparov Event Horizon for Object Perception & Handling
- 4.3.23 Scientific Method and Data Science Models

- 4.3.24 AI Bodies and Brains
- 4.3.25 Interpreting Hofstadter's Gap: AI, Music, Math, Language
- 4.3.26 Modularizing Problem Space for AI
- 4.3.27 Questions as Objects in AI Object Handling
- 4.3.28 AI, Biology, & Types of Intelligence
- 4.3.29 CS, STEM, law, natural-law
- 4.3.30.1 Architectural Learning, Developmental-Landscape  
Hypervolumes, & Empirical Task-Trees: Participation
- 4.3.30.2 Calculating Tea for AI:  
Advocating for Architectural Learning
- 4.3.31 Subroutine Stacking Measure: AI Learning
- 4.3.32 AI ALU Corpus Callosum
- 4.3.33 The Vortex: Gender and AI's Scylla and Charybdis
- 4.3.34 Modeling Participant Architectural Learning  
in Five Trees Plus Mindstate
- 4.3.35 Biology, Psychology, Math: AI Broad or AI Narrow
- 4.3.36 Modular Architectural Learning: Jellyfish Spider-Crab AI
- 4.3.37 Ant Tasks vs AI-Agent Tasks

#### **Appendices:**

- Appendix #: Recommended Reading & Extended Reading
- Appendix #: Definition Behavior Studies Mnemonic
- Appendix #: Expanded Introduction
- Appendix #: Expanded Part 1
- Appendix #: Expanded Part 2
- Appendix #: Expanded Part 3
- Appendix #: Expanded Part 4

## **A Narrative Introduction in Two Parts:**

**Introduction Part 1. Chess in Blade Runner**

**Introduction Part 2. Defining AI Goals and 'Objects'**

### **Introduction Part 1**

#### **Chess in Blade Runner: AI in 1968 vs. AI in 2023**

When the film 'Blade Runner' was released in 1982 (based on a PKD book from 1968) the idea of a narrow, un-thinking, single-purpose, chess-AI was not something that most people thought was even possible.

In 1982 chess was not seen as a narrow set of math problems that hardware and/or software could brute force well enough to defeat a human world champion: An ability to play chess was a broad measure of the human intellect, harkening back to the European chess cafe's of the enlightenment era when Benjamin Franklin would enjoy being trounced at chess amid France's philosophical discussions of the age covering all areas of STEM, arts, humanities, culture, politics, the marvels of the natural world, and more. Chess had become symbolically entwined with the expansive and romantic view of humanity and human-ness, and of the world itself as a vast interconnected and promising



realm, a "revolution" against the backdrop of oppressive doctrine that overwhelmingly and stiflingly defined people and the world as merely a dull evil destined to be destroyed by an unimaginable 'goodness' from an alternate dimension. That there was character and depth and meaning and discovery and imagination in this world, in humans, and in nature, was a breakthrough epiphany that not so long before had been so politically dangerous and daring that people like Geordano Bruno were literally burned at the stake in public to let it be known what happens to little people who do not know their place a feudal prison of mind and body. And, if undeservingly, chess was mythologized and embraced as part and parcel of this humanist-naturalist empire of the science-fiction and fantasy imagination. While this was not the same in every country or region, even in the USA where chess has not been followed as closely as in Europe, the same symbolism was often still there. And for many people, the ever present (yet also invisible) Claud Shannon and Alan Turing being rare exceptions, the romantic symbolism of chess completely overtook the idea that chess was a mechanical game with rules. It was often stated that for a computer to play chess it must master the whole human mind and realm, and that if a computer could someday beat a strong human player, that this all-wise machine would be poised to do everything that any person could do, only perfectly, and then take over the world!

Since 1968 (and the eras that came before) both the field of AI and the game of chess have fallen into cynical, modernist malaise which has probably gone a bit too far in the opposite pendulum-swing direction: the world again is "merely" a narrow bitter zero-sum or negative-sum game. Yet, chess still seems to be at the center of how people view the world...no mind in chess: no mind in the world, no mind in AI.

And perhaps in a similar, parallel, or rhyming way, Science Fiction has also contracted into a 'hard science fiction' that is not so interested in exploring the strange depths and dimensions of consciousness and reality. Compare the works of Theodor Sturgeon, and Philip K. Dick (and even Robert A. Heinlein) to later titles.

But during the 1980's when Roy Batty, a machine, a synthetic, mastered chess! Why that meant having a renaissance enlightenment within an AI-mind. And just so, along these lines both the film 'Blade Runner' and the book (though perhaps more-so other PKD books, such as 'Flow My Tears The Policeman Said') emphasized the old-world depth of cultured life. The chess board that Roy Batty learned to play on was in the film a traditional old English 'Birds of England' chess set, eccentrically mismatched from the standard bird pieces, as though it was (and maybe the set picked for the film actually was!) a mismatched heirloom from the 1800's spotted in someone's Aunt's countryside farm house in rural England or France, looking the part coming from some past forgotten time of whimsical and mysterious minds and old imaginings of the natural world and of natural philosophy. In many PKD books the characters and plots hinge, though they take place in a technologically advanced future, on seeking an intellectual and emotional appreciation of historical arts: of music, of ceramics, philosophy,

and theology. The characters are often so obsessed with art and the natural world that they barely care about the high technology.

This element of chess in the film's story may even have been intended to be a key part of character development in the plot, a key revelation and turning point to make things plain yet astounding for those watching the film. Though Americans have shared the view that chess represents a roundly superior mind, the English Ridley Scott, with his eccentric old English chess set, was the one to put chess centrally into the short-story format of the cinema, where every visual scene must operate economically on myriad levels and tell layers of stories on many dimensions simultaneously. As the story goes, Roy Batty learns and masters the game of chess (on that antique naturalist European chess set, though the story takes place in Los Angeles). Roy Batty then defeats his own creator, a human genius, at a game of chess. After this unimaginable victory his creator likens Roy to the flame of a candle, saying "and you have burned so very very brightly, Roy." After this Roy Batty becomes poetic, makes peace with his own mortality (against which he struggled throughout the story), and takes pity on the film's (perhaps human) protagonist Deckard in a miraculous change of heart, saving his own adversary from certain death.

Those after 2019 may have trouble following this path of character development. How did this synthetic-man burn "so very very brightly"? What did he, an android, care what happened to a human? But to those from an earlier era, for a machine to have mastered the embodiment of the enlightenment, to have mastered what it meant for mortals to think, strategize, and imagine, would be an indication of some great and subtle internal awakening and transformation. That a machine could through dextrous intellectual skill, embody any renaissance talent, could, like Sherlock Holmes, unwind any situation, plot, device, or unfold any mystery at a glance, and could engender all ethical and humanistic apprehension: the depths of the heart, the mysteries of symphonies and operas; the rational puzzle-solving mind was seen as the essence of all existence; indeed since Laplace people said that if sharp enough a mind could know with absolute certainty the entire history and future of everything in the entire universe and know intimately the mind of the creator of the universe: Such a feat would be indeed a bright promethean flame, and such a flame of mind was just what chess symbolized.

## Deep Blue & Blade Runner Chess

In real life a machine (of sorts) did defeat the world chess champion, arguably, in a match in 1997.

It is very interesting to compare the machine that played against Gary Kasparov to the machine-android in the film Blade Runner, and to other forms of AI that existed in 2019.

The actual Deep Blue system was never made public which is another twist in the very labyrinthine story. Many consider this a slight to the history of science

and the many people around the world including Gary Kasparaov who had worked hard to cultivate an international computer chess software community since the 1980's. So what is known is largely indirect information that was allowed past the extreme legal secrecy contracts that IBM surrounded its program with.

To greatly over-simplify this here (it is a huge sprawling, interesting, topic on which many books and articles have been written), let's look at types of approaches that could have been taken and then pick out which systems were deployed to make Deep Blue play chess.

Here are optional areas of approach. (Which did Deep Blue actually take?)

1. Douglass Hofesteer type 'conceptual understanding,' a machine with self reflecting consciousness, feelings, beliefs, attitudes, artistic opinions, etc.
2. a Newell, Shaw, Simon 'symbolic' & linguistic human-type cognitive reasoning system (which no one has found and may be a reification)
3. Analogy, pattern, and space apprehension.
4. Frequentist Statistical Learning & Parametric Machine learning (Like fitting a cartesian X Y bar graph to a set of points to get the slope of a general line.)
5. 'Sub-symbolic' Neural Networks and 'Concept-learning' Embedding Vectors
6. Decision Trees and Bayesian Statistics
7. Pre-written expert system decision procedures (for first half of game 6. (openings) to use generally.
8. Genetic algorithms and reinforcement learning. ('Evolution' by selection pressure.)
9. unsupervised machine learning
10. supervised machine learning
11. Human manual instructions for specific cases
12. AI-self-boot-strapping: procedures for how to start a process (game)
13. human-force-feeding-boot-strapping: a human manually forced the AI to start a game according to a panel of human experts
14. disinformation procedures to psychologically attack and cause distress in the user
15. Human manual inputs and changes to the system at the last minute, including human choices for ending the game (draw, resign, accept draw, etc.)
16. A Human Spy network to feed in 'cheat' information manually.
17. A robotic body to move the chess pieces for itself
18. Human intersession to move the pieces for the AI (often making mistakes)
19. general purpose hardware and software that could run on various hardware
20. specific hardware and some software built for one operation (not cross-platform standard software)
21. automated interaction between components
22. Bayesian Learning

Out of these various options Deep Blue used:

7. Pre-written expert system decision procedures (for first half of game (openings) to use generally.
11. Human manual instructions for specific cases

13. human-force-feeding-boot-strapping: a human manually forced the AI to start a game according to a panel of human experts
14. disinformation procedures to psychologically attack and cause distress in the user
15. Human manual inputs and changes to the system at the last minute, including human choices for ending the game (draw, resign, accept draw, etc.)
16. A Human Spy network to feed in 'cheat' information manually.
18. Human intersession to move the pieces for the AI (often making mistakes)
20. specific custom hardware and some software for one type of tree-search operation (not cross-platform standard software)

(Note: There may be some parallels between Deep Blue as a milestone in "AI" and ENIAC as a milestone in "programmable" computing. Both Deep Blue and ENIAC had very significant manual-human processes in how they worked. This manual-human-connection may both make Deep Blue and ENIAC somewhat awkward examples of non-human later technologies (e.g. a 'program' for ENIAC was people (brilliant women programmers) running around with wires and manually experimenting with dials, which is not how we think of software programming today (though there are very direct connections such as 'Breakpoint' setting in debugging), and perhaps this also emphasizes the enduring pattern of computer science and AI being fundamentally part of H.sapiens-human, biological, and ecological processes, not remote abstractions operating in some alternate reality entirely on their own.)

### **An AI Chess Participant**

At the time it was enough to say 'Deep Blue won' if it did a few of the steps of playing 'on its own.' But it is interesting to think about what would be needed for an AI-robot to participate in and win a chess match (or tournament) with no human intervention or support.

If you had simply put Deep Blue and Gary Kasparov in a room together, with no human intersession to direct or correct or supplement Deep Blue, then Deep Blue would not even have been able to play chess at all on many levels: it could not start, it could not decide how to open the game, it could not stop a game, it could not move pieces on the board, it could not see the board, it could not reboot after frequent crashes: Deep Blue's distant super-computer constantly crashed and needed rebooting, resetting, and preening by an army of engineers (and chess experts) just to get through and complete each single game.

This gives us an excellent real-world example of what we think of as a whole AI system or parts of an AI system. Deep Blue did a few very specific parts, it performed a few 'project-roles' (or assigned sub-tasks), out of a whole set of tasks in a larger project. In some cases it will be clear what an AI should do to be 'independent,' but there is likely a lot of gray area too.

For example, it would be very interesting to set up a chess match where it was human vs. computer, and the computer (AI-robot) had to complete the entire match 100% on its own. Such skill-ability items might include:

- entering the room
- supplying power, like a battery (as the humans bring 'fuel,' as Kasparov termed it "bananas and chocolate")
- starting and ending the game
- seeing the board
- making moves (moving chess pieces)
- using the game-clock
- being able to suggest and except a draw or resignation
- finding out and getting to the correct time and place for a game when the event schedule may be set at or change at short notice

and probably with the same rules that humans have to use: no help, no internet connection, no phone, etc. basically in a faraday cage with candles for light and older technology for everything, and possibly traditional practices such as the post-game discussion where the two players talk about the game when the match is done.

Could a robot do this in 2023? We are getting much closer...but most likely not yet.

The power-supply issue itself is a very interesting part of this. There is a fundamental relationship between the available power-supply and the depth and speed of move-computation. Many chess programs will soak up whatever resources are available, throttling itself based on what energy and computational resources it can get. If a computer has 'unlimited [electrical] power', and funding you could put as many parallel processor cores into the computer as you could fit and end up with essentially a super-computer focused just on the chess game (very expensive, and very not-portable). But having a portable 'walking' robot that can last 7+ hours to finish a single game (let alone a six-game match or a whole tournament), and a realistic budget of time and resources, puts some interesting trade-offs into the design. And this is not necessarily unprecedented or unreasonable. Professional chess players use elaborate consideration and preparation to pace and regulate their own resources to stay sharp when they need to be. In fact there is a lot of attention and some fuss over the details of how chess players are allowed to 'draw' on whole games just so they can focus on a next game that matters more for their overall tournament score (or some other item), because humans do not have enough stamina to put all resources into every game. There are probably parallels in athletics such as the Olympics. Arguably a big part of being a professional chess player is the ability to manage all the processes of focus and stress and game preparation etc., which goes well beyond playing one game under ideal conditions. It would make sense to construct a match where a chess AI is expected to 'self-manage' resources in the same way.

## Introduction Part 2. Defining AI Goals and 'Objects'

### *Goals & Project-Objects in AI OS (Operating Systems) & Architecture Problem-Space(s)*

**What is needed for the AI systems we want to deploy?**

- What is the goal?
- What can't AI systems do yet in a context of that goal?
- What are the specific sub-skill, sub-part, needs for meeting the overall goal?

### **The Woz Test: Project Participant & Portable**

Goal: Project Participation

Sub-Parts of Goal:

- communicate (as part of task completion)
- plan (as part of task completion)
- develop, complete, assign, and close-out, tasks alone and with other participants.
- be responsible for tasks
- follow and maintain the project schedule
- handle 'project-objects'
- assess health and feasibility of project
- participate in project planning
- participate in iterative project review
- give and receive feedback
- make and run evaluations and tests
- initiate events & actions
- report and document (including external project-object data) etc.

Having looked at Deep Blue as being very far from the ability to walk into a chess tournament and walk out a week later having played with other players on chess boards (winning being somewhat beside the point), let's look at what some people call 'The Woz Test' for AI, which is (with lots of variations) the ability to make a cup of coffee.

I do not mean to get hung up on this or that detail, but rather focus on something like 'making coffee in an office with team-members' as an example of being able to participate in a project with other participants.

### **Before vs. After ChatGPT & Large Language Models (foundation models)**

#### **What was needed before LLM & ChatGPT?**

Still looking at the Woz-Office-Coffee-AI-Robot goal: if we can climb into our way-back-machine, back into that time-capsule students buried outside the library way back in the ancient times before chatGPT (perhaps in Blade

Runner's cathedral year of 2019), if we were taking on the Woz challenge then, what might our list of goals and tasks be?

### **AI OS & Architecture:**

#### **Systems needed for Woz-AI-Coffee-Bot: before 2023 & LLM & ChatGPT**

(some examples)

- networked component-AI (connectable)
- AI-OS for single components
- AI-OS for networked-multi-components
- whole-AI multi-component architecture
- external object handling for projects
- external object handling for component-AI
- project objects (in object relationship space)
- external-project-object database (of whatever type(s))
- general vs. deployment: scope & resource policy
- (project) object-relationship-space map

### **After:**

#### **What was 'new' after the revolution of LLM & ChatGPT?**

When Large Language Models and OpenAI's ChatGPT opened the revolution and turned world side down, and changed the game, and many other not entirely clear phrases, what exactly had changed in this Woz-Bot context? What was introduced? What was possible that had not been possible before?

#### **After 2023 & LLM & ChatGPT: What is new?**

*new item = internal object handling*

So...hm. There is one thing on the list of world-changing-raptures (*new item = internal object handling*). And the one thing that is different was not even on the original list of needed things. Well, maybe this new emergence and discovery changes 'the game' as they say and changes what is needed? Let's see. What was then *still* needed to be done (or still needs to be done now) after LLM and ChatGPT? Did this unexpected new ability replace or fulfill older needs?

#### **What was still needed after the 2023 revolution of LLM & ChatGPT?**

List of what AI systems need to cover: after 2023 & LLM & ChatGPT  
...exactly the same list as before 2023

Depending on availability, portability, and resource cost, a Large Language Model could conceivably help by adding the ability to 'internally' handle 'objects.' However, that is largely an 'under the hood' design-choice detail of 'how' the AI works (under the hood). The overall project requirements are:

- external object handling for projects
- external object handling for component-AI

Both of these have an uncertain connection to what ChatGPT can do internally. I am optimistic that LLM and internal handling will help greatly, but case by case the actual system architecture needs may be difficult to solve.

This is not to dismiss out of hand the very practical ability to do internal project object handling, the idea is to point out a few things (and immunize ourselves to shifts in hyperbolic rhetoric from 'nothing is possible!' to 'everything has already been done!):

- To have a sense of the landscape of AI Architecture & OS Systems, for example the list of needs we came up with is generally not discussed at all in specific model technology discussions that are focused just on usually single-purpose models that do one sub-component function.

- We should more clearly understand the abilities of internal object handling case by base for applications. For example, have you ever heard people talk about, or give test results, for internal project-object handling abilities of Large Language Models?

- Limitations such as Kasparov Event Horizons, which affect internal object handling more so than external project data.

-

The goal of this paper is to contribute to better overall discussions, planning, coordination, etc. around the larger landscape of **AI OS & Architecture**. For example, let's compare two 'bot's that at first may not seem very different. See here: <https://www.youtube.com/watch?v=rnIgnS8Susg> On the ball, as usual, the wonderful Khan Academy has implemented a (for example) math skill helper bot, like a tutor for students who need help. For the specific deployment case of a discrete-topic bot who lives on a website and answers logic questions:

*Requirements:*

- *Answer good logic questions.*
- *Refuse bad questions and comments politely.*

OpenAI's Large Language Model technology solved the needs and deployed the solution in one fell swoop: task done! Problem solved. Project Completed. Not only is Khan Academy closer, they are done: system deployed and working.

But for a Woz-Bot deployed in your office, we are in many ways no closer at all, even one that lives in the cloud and used a networked coffee maker may be far away.

Something to keep an eye on is whether robots in warehouses such as Amazon's find ways to use Large Language Model technology, if there is warehouse AI 'participation', or if, even after the ChatGPT revolution, there is no intersection, interconnection, and the only things that AI can do in a warehouse still are (usually, without crashing) moving a cart along a painfully obviously painted line on the floor exactly as told to do so. That may be a terrible example, but question is: where will various AI technologies be applicable and not applicable in the bigger picture of AI OS & Architectures.



**A minimal illustration:** "Did I just show you a picture?"

In case it is helpful, let's walk through a very minimal example that may, despite being tiny, put some tangible detail on some of the abstract design factors that we are trying to plan out.

A common rule of thumb before ChatGPT showed internal object handling was:

*"Most AI are very good at producing a correct output in one specific task, but not very good at doing anything with that information."*

An interesting question (that I have no idea how to predict far into the future) is where is this totally changed by OpenAI's LLM, and where is it largely unchanged? Time will tell.

Let's say you show an AI two cat pictures, which we will say is asking the AI what those pictures are, and the AI correctly identifies both pictures as cats. So far, so good! You gave the AI two pictures as inputs, effectively asking it to identify those pictures, and it output two answers. Let's do a quick tally.

*Input = 2 pictures you input into the AI: picture\_1, and picture\_2  
Output = 2 answers that the AI gave back to you: "cat", and 'cat'*

Now, what if you asked the AI some simple questions:

|   |       |
|---|-------|
| "How many pictures did I just show you?"      | [Two] |
| "Did I show you a picture of a cat?"          | [Yes] |
| "Were the answers you gave me both the same?" | [Yes] |

In short, most AI will not be able to answer these questions correctly.

Notice, here you are essentially asking the AI about its own 'state' (people will bicker about the semantics, but something like that). Not only does the AI have this information, this is the only information the AI does have. But, yet, the AI still cannot answer.

We will look at two types of reasons why an AI might not be able to answer your simple questions about information that it does have about its own state. One of them is a bit strange or funny, and may seem like a joke, but it is still important, and interestingly still relevant even after the emergence of OpenAI's ChatGPT and Large Language Models.

Reason 1.

The AI does not have the ability to handle 'objects.' It has data but it cannot granularly separate, distinguish, and handle individual 'object' elements with properties such as how that object relates to other objects (or "object-relationships").

Reason 2.

The AI only takes image-files as input and cannot answer Natural Language word-questions. All you can input is a picture, and all it can output is an identification output. This might sound ridiculous to say, but it may actually be very important.

These two reasons can be taken together to help puzzle out more about the operational space of AI, or the realm of Architectures and Operating Systems, and also to think critically about what we read about AI.

Regarding the Reason 1, it does not matter per se how the AI is able to handle objects. In the past this has been a major area of dispute and speculation in AI research, and likely will continue to be important in various ways. For example, on the level of AI system architecture, in some cases it likely does not matter at all how the object is handled so long as it is (is the box taped from the right side or the left side?: the result is all that matters). But once you go beyond that level to 'external' project-objects that must pass between components and be logged and read and shared and updated etc., here the details are very important.

A possible example here, hopefully not a terrible example, is: imagine two different scenarios for the Khan Academy Chat-bot.

Scenario 1: The Khan-Bot answers logic questions in a browser. There really are no 'external project objects' here. Everything is forgotten. There is no project-participation. No real world object inventories or records are changed. It doesn't even really matter if what the AI says is nonsense.

Scenario 1: The Khan-Bot is on the board of trustees for a school and on the committee to design and implement the next year's math curriculum, and report on effectiveness at the end of the year. From an AI OS & Architecture perspective, this is a very different system. In this case there are many 'external project objects' that must be dealt with along with other participants in the project. The output of the AI is not just something it can generate and delete with no world consequences. Everything is still made of 'data' (this is not physical warehouse item movement) but the data are now 'external' real world 'objects' in a project-space, not just an internal arbitrary amnesiac sandbox that evaporates the same way each time. There is now a massive system of data moving through many components and between many participants, there are schedules, there are deliverables, there are agreements and disputes, there are evaluations, etc. etc., and no 'internal' ability no matter how profound will automatically 'solve' this entire multi-participant process. And this kind of leads us to the second reason mentioned above.

The second reason holds another set of keys: Components, and communication between components, and projects, and communication about project-objects.

### **'Internal vs. External'**

Many researchers in the past speculated that internal object handling was impossible...which turned out not to be the case as illustrated by OpenAi's Chat GPT & Large Language Models. To attempt to illustrate this, let's invoke the idea of a project-object-database (How exciting!). To use the nicely tiny example of cat pictures above, such a project-object-database would be a manageably small affair, a table of information where you could look up information about the inputs and outputs. Simply by reading this table you could answer the questions you asked. Now, while it is possible that some AI systems might benefit from using such a table to 'do something with' the output information that it so expertly output, OpenAi's Chat GPT & Large Language Models have demonstrated that no such literal database is needed in all cases for internal processing. Indeed, OpenAi's Chat GPT & Large Language Models do far better with what is called 'sub-symbolic' internal processing than any past 'symbolic' AI was able to do at answering questions from a database. However! There is a big difference between having or not having an internal project-object-database for purely internal processing and handling of objects within a component, and on the other hand there being an external-project-object-database for external project data that is managed by multiple project participants.

Perhaps using the fullest vague extent of the general definition of "database" as 'a collection of information,' the exact form or forms of these external project-object-data does not matter (so long as problems don't harm the project, such as data-loss from failed storage and no backups). It might be a literal single mega-database maintained by the AI for reference and sharing of project data, or it might be a 'proverbial' database that is merely the set of many other databases and sources of information that are 'collected' by the project. Either way, the AI Architecture must interface with and interact with that external ~database. And this likely means many back and forth translations between symbolic and sub-symbolic processes (even if everything 'internal' to the AI is sub-symbolic), the project as a whole will not (so far as I imperfectly predict) be entirely sub-symbolic.

### **Mix and Match and Generalize**

While I may be missing something, it seems peculiar that so many books on AI use the example of an image-processing AI to argue that AI does not (and some argue 'cannot') understanding language concepts: for example the classic notion that an I can classify a picture as a cat but it does not understand linguistic concepts about "cat."

There are probably many issues being mixed together here, and perhaps that is part of the problem. In some cases the goal may be to make the case against alarmist exaggerations that picture-classifying AI, kind of like claims about like Big-Blue-Chess, will suddenly be able to do everything humans can do but better and then 'take over the world.' So in that case perhaps the focus is the rather obvious mismatch that it makes no sense to ask a single-purpose narrow-AI to do a some other task. Picture-input-only AI only takes pictures as input. Sound-file-input-only-AI only accepts sound files. Natural-language-input-AI may only accept letter and number characters as

input. So there is no danger that a picture-input-only AI that has only been trained on cat images will suddenly start tasting wine, and writing essays, predicting tomorrow's barometric pressure, and controlling robotic arms, etc. Let's say this case is fine and set this aside.

In other cases people do seem to be moving in the opposite direction and mix-matching language concepts with labeled photos. For example, the standard paradigm statement that an AI can classify a picture as a cat but the AI does not 'understanding' and discuss the language-concept of 'cat-ness' because AI lacks the (largely undefined) auto-instant-general-transfer-learning-intelligence-conciseness special sauce that h.sapiens-humans (according to them) have. It is possible that this is something of a semantic disagreement, and the people making this argument would say that they are not talking about 'Language-concepts' about cats and animals and cat behavior etc. etc., they may claim they are talking about 'general' concepts such that the concepts could be applied equally to images.

I think this raises a number of fascinating and likely at least as of 2023 not yet resolved issues and questions. This raises AI-ML model and training questions. This raises h.sapiens-human brain's structure questions. This raises questions about linguistics and perhaps the psycho-linguistics of how people view language (or how language is invisible to people).

Whether or not there is such a thing as a non-linguistic concept of cat-ness may be a rather philosophical question, and it may not be clear in 2023 whether it is useful or even definable. It may turn out to be, but I would say it is not yet clear. Especially perhaps since the context of these claims is that the now clearly existing Large Language Models with object-handling abilities are not possible...which is of course false.

For example, there are several very real and closely related sets of issues, but it is not clear how they relate to the original claims (which in various ways have already demonstrated themselves to be bogus in their completely and clearly wrong predictions, e.g. that LLM internal object handling will never happen because either that or any sub-symbolic object handling is impossible in principle.) For example, keeping the topic of 'image' + 'language' and focusing on very real questions of how AI can, and perhaps how the h.sapiens-human brain does, connect largely separate systems and components that process images with components that process language-concepts.

I would also like to point out the combined ideas that

A. people don't understand what language is, and completely consistent with that is

B. that 'language space' operates in ways that people don't yet understand (should not be surprising there). In the past people micromanaged what they wanted 'symbol' and 'language' to mean and do in hand-crafted AI decision systems (so-called 'symbolic AI'), which, perhaps not surprisingly, did not work well outside of a small number of very specific narrow finite cases.

'Words' do not simply equal 'language' or 'linguistic concepts'

People, especially in the west, often make the blanket assumption that the world and language are made of words and concepts which are the essentially the same, and the way they use language is often invisible to them. But a major issue that quickly comes up when doing Natural Language Processing (whether or not it is called 'AI'), is that 'words' and concepts are not simply the same and neither are simple to define. OpenAI's Large Language Model and ChatGPT is not (or not only) trained with human-micro-managed sets of pre-defined 'words.' We like to think of concepts and words as being clearly equivalent, but in reality a concept will correspond to a large and fuzzy set of real world language characters and symbols. (Note: One could ask and no doubt some will both out of curiosity and for 'devils advocate' trolling': how do we know concepts exist? Here I would direct you to Francois Challet's Deep Learning in Python where you will find hundreds of pages explaining how to use today's analytical statistical and hypothesis testing tools to make and test that question: welcome to what AI-ML model creation is and does!)

Another part of this may be the rather inexplicable statement that AI does not 'understand concepts.' Perhaps an AI's concept is insufficient, or unlike human concepts, or maybe a given person has a semantics-lexicon issue with the use of the word 'concept' but by any reasonable common sense description, concepts are exactly what subsymbolic deep learning models.

To put these parts together, when a sub-symbolic deep learning AI is trained on language characters, it is literally constructing linguistic constructs of things like 'cat' and anything else that can be described in language, but not based on words.

And so there is a bit of a 'language problem' here. Also: we are using language to have this discussion of concepts and it is not clear (to me at least) how we would have this discussion without language.

### **Taking a Step Back: Languages, Images, & Concepts**

The fact that the first AI to be able to do object handling of any kind, let alone internal-object handling, came from Large Language Models trained on character-gram inputs of not 'words' (pre-defined by people) but raw streams of characters, may not be a coincidence. That we are having this discussion about language, images and concepts, using language (and not images) may not be a coincidence. That the only biological species to discuss the concept of cat-ness is a language-using species may not be a coincidence. That the first AI to be able to participate in a discussion with a human about catness was a language-only AI may not be a coincidence. Can we separate concepts from language? What would that mean, and why would we want to do so? Is there some function in doing so?

Is there such a thing as a purely image-based "concept" of catness that specifically excludes "language" and "language concepts" (but also includes "language concepts")? That question sounds like self-contradictory rubbish to me.

We should be thinking about AI in a larger context of AI Architecture and AI OS that includes projects, participation, roles and tasks, handling objects, internal and external objects, projects objects in a projects space and multi-participant space (and multi-component space). Questions such as how does the h.sapiens-human brain handling both images and language concepts, is perhaps important and useful in various ways, for example hybridizing a combined h.sapiens-human+AI things (for example as a treatment for stroke victims with vision and or language processing brain injuries, or just enhancements so people can be smart enough to manage resources without destroying everything). How will very often separate single-purpose components of AI work together in and AI? How good or bad is the ability to bridge a very sophisticated language model with quite possibly a much less sophisticated image model? (Note: I did experiment with asking ChaptGPT to create ascii art, and it seemed to have almost no ability to make a coherent picture). How will multiple AI participants work together? How will AI participants work together with h.sapeint-humans partiipans in a projects-space? How will image and language concpets work together across projects-spaces using dignital and AI tools? These are practidal quetions.

#### **Factors:**

- Internal Skills vs. External Skills
- Internal Data(base) vs. External Data(base)
- Single component vs. networked components
- Specific-deployment vs. general
- Types of signal/data
- Single signal/data-type vs. multi signal/data-type
- Project-space skills needed

#### **The 'Ghost' of Deep Blue**

As a note, being able to make coffee may include being portable enough to do so. The 'ghost' of Deep Blue may indefinitely hang over AI, where there are super-abilities but only if you have a connection to a private supercomputer and virtually unlimited resources. Perhaps in smart-cities there will be a gray-areas of easy supercomputer links within an urban zone; In 2023 is it simply too early to tell what will or won't be issues for different kinds of deployed AI systems. If one is planning in general, there may be many kinds of options such as portability or wifi-access to include. If one is planning very specifically, they may need to clearly identify the niche of that deployment. And then maybe the world will change overnight anyway.

## **Part 1: Object Relationship Space Framework**

- 1.1 Example General Object Relationship Space List
- 1.2 Many lists in One
- 1.3 Networked-AI Components

## 1. General Object Relationship Spaces (Example List)

### Object Relationship Spaces:

- AI-ML-DS Model Space(s)
  - regression statistics
  - classification statistics
  - sentiment etc. analysis
  - convolutional
  - n-grams
  - bayesian
  - deep learning
  - transformers
  - Generative Adversarial Models
  - genetic algorithms
  - reinforcement models
  - etc. (many more, and list will grow in time)
- Assignment(Role, Task) & Delegation Space(s) (Projets)
- should, ought space(s)
- Body Space(s) (self-maintenance)
  - hardware
  - low level OS
  - AI-OS
  - containers, virtual spaces
  - networks
  - personas
  - NLP engine
- Categories of Types of Systems Spaces(s) (Generalized STEM)
  - Abstract-Logic Space(s) [always hypothetical]
  - Statistical-bridge-between-physics and math-logic space(s)
  - One-tree Physical Space(s)
  - Dynamical & Fractal near off the one-tree Space(s)
  - Management of categories of types of systems
- Code Execution Space(s)
- Component Network (project-context Networked-Intelligence):
  - low level components (internal)
  - high level components (external)
- Confidence, Probability, Noise-level, level of guessing.
- Documentation Space(s)
- Essence/style/sentiment Space(s)
- Explanation of Process Space(s) [model explainability]
- Feedback and Testing Space(s)
- Gamification Space(s)
- Hypothetical & possible Spaces(s)

- Instructions Procedures Space(s)
- Low Level Files Space(s)
- Network Space(s)
  - firewall
  - servers
- ?NLP Space / Natural Language Space(s)
- Object Attribute Database Space(s)
- Quarantine Space(s)
- Plans, Flags, Reminders and Notifications Space(s)
  - check when getting signal
- Policy Space(s)
- Project Management & General System Best Practice Space(s)
  - Schedules
  - Roles
  - Tasks
  - Documentation
- Project/agile space(s)
  - specific project data
- ?Question Space(s)
- ?reality space(s)
- Recycle/Trash/Disposal Space
- Reference / Library Space(s)
  - private notes
  - private data
  - external data
  - storage/archive
- Sandboxes Spaces(s)
- Security
- Sentiment, Appropriateness
- System-1 & System-2 Space(s)
- System Fitness, Collapse, Ethics Space(s)
- STEM space (one-tree?)
- Taxonomy Space(s)
- Time & Schedule Space(s)
- ?Translation/Conversion Space(s)
- ?
- New Spaces made by the AI
- statistics-to-one-tree space
- near-off-the-one-tree space

Note: A list you will use for a project will be a list for that project. This is a general example list for illustration purposes. It is unlikely that there will be a portable list of every possible part of every possible project in the universe. Figure out what your context is, what you need, and what your schedule is. (Again, see the 'general vs. deployment' issue.)

## 1.2 Many lists in One



The above list is abstract and can be contextualized in many ways:

e.g.

- Skills/Abilities
- Objects
- Types of Objects
- Relationships between Types of Objects
- The problem-spaces for relationships between types of objects.
- Tests for Object-Spaces & Object-Relationships
- Project Goals
- Components (low level and high level)
  
- Internal vs. External
- Networked vs. Single-State
- Project Scale
- Project Roles

*See notes on breaking this up in different ways in the appendix [<here>](#)*

### **1.3 Networked-AI Components**

Another context of AI, [which is also 'things on the list'](#) is components that are networked together.

A network may extend in an in-ward direction. For example an AI that is not a single-blob 'end-to-end' model that wraps all functions together, then those separate functions are done by separate internal components..

Sometimes a functioning AI-bot will need to have multiple collaborating AI parts, which are able to work together within a network of components, and different levels networking with other components and other AI-bots.

Usually a project overall has many participants, and if the AI is more broadly participating beyond being like a silent screwdriver used by one human then the network may extend outside of the AI in question.

This topic will also be an ongoing theme, for example in terms of design decisions and trade-offs for a given project: how much to use one-blob, and how much to use a hybrid ensemble composite mix network of components.

#### **Low Level Components:**

- computer-vision components

- NLP components
- audio components
- general system-1 component
- general top layer system-2 component
- generation component
- internet/intranet/network component
- EM spectrum component
- image-to-text component
- audio-to-text component
- file-to-text component
- server and firewall components
- 'eyes and ears' components
- 'arms and legs' components
- automated documentation component
- container and virtualization management components
- low level data storage & database navigation & management components

### **Higher Level Components:**

- Schedule Management
- Project Management
- Self-Status Management
- General System Health & Security Management
- DS AI ML model manager:
  - manage tasks with known data and models
  - understand new problems (types of data, etc.)
  - match known models to new problems
  - modify models
  - make new models
  - manage resources (cpu, gpu, local, cloud)
- managing feedback:
  - interpreting feedback input
  - crafting feedback output
- persona manager & PR/HR problems
- high level data storage & database navigation & management components

### **Specific Processing Units:**

As common computer architectures have an ALU (Arithmetic Logic Unit) and a CPU (a Central Processing Unit) and now often a GPU (for Graphics Processing and matrix operations), a sometimes even a TPU (Tensor Processing unit), will future AI systems and AI operating-systems, have different processing-units that they feed different kinds of low-level problems into?

(Note, in a later section there is an example adding additional "steps" of "levels" on top of a simplest function. Depending on the project, there are many contexts you could use to describe what in the end seem to be all items in this object list.)

### 1.2.3 Organismic whole-body parts of AI

Narrative: we think of the human mind-body or the human brain as being 'just one thing' because of how we experience it, kind of as a user-interface experience, but it really does have different parts (if any one part stops working various things go wrong).

Even the brain has sub-parts (of often they have sub-sub-parts). E.g. While it is derigur to poopoo any connection between AI Neural nets and the brain (no silliness. As Patric Stewarts once said: "We are not here to have fun!") The human brain really does do 'edge-detection' as a specific sub-feature of breaking down "computer-vision" problems for the biological brain.

## Part 2: Using The Framework

### Part 2: Using The Framework

- 2.1 Examples: AI Skills, Comparing Three Chatbots
- 2.2 Examples: AI Skills Mapped to Object Spaces
- 2.3 Adding Levels, Adding Steps, Adding Objects
- 2.4 Hybrid/cross-model skills for AI
- 2.5 Heuristic & Pseudocode for AI Management with Object Spaces
- 2.6 Modularity, Scale, and Networks
  - whole ai, networked ai, ecosystems

### 2.1 Examples: AI Skills, Comparing Three Chatbots

*Systematizing how AI uses objects*

The same five questions will be asked of three leading chatbots of 2023, who were designed to operate with goals, and not surprisingly do not all perform at the same level when compared across different tests.

| Questions   | Kuki AI   | Wolfram Alpha  | ChatGPT 2023.3   |
|---|---|--|--|
| <p>1. If <math>A = B</math>, and <math>B = C</math>, Does <math>A = C</math>?</p> <p>[Logic, Math]</p>                  | <p>C. The C is for computer.</p> <p><i>Note: She will also give this answer if you replace C in the question with D. Apparently she likes alphabet games. Adorable.</i></p> | <p><math>B + B = C</math></p>  | <p>Yes, if A is equal to B, and B is equal to C, then A is also equal to C. This is a fundamental property of equality known as the transitive property. In mathematical notation, the transitive property of equality can be expressed as follows:<br/> If <math>A = B</math> and <math>B = C</math>, then <math>A = C</math>.<br/> For example, if we know that <math>A = 2</math>, <math>B = 2</math>, and <math>C = 2</math>, then we can say that <math>A = C</math> because <math>A = B</math> and <math>B = C</math>.<br/> Therefore, <math>A = C = 2</math>.</p>   |
| <p>2. Can you look up an image of a rose?</p> <p>[Abilities, media files, talking about self, internet connections]</p> | <p>[rose bud pic shown]</p> <p>I hope you like this picture.</p>  | <p>Assuming "rose" is a plane curve   Use as a plant or a color or a neighborhood or an administrative division instead interpretation rose curve   Result Equations</p> <p>Note: Wolfram Alpha can generate math graphs and charts.</p> | <p>As an AI language model, I don't have the ability to browse the internet or display images directly. However, I can provide you with a text-based description of a rose.</p> <p>A rose is a flowering plant that belongs to the Rosaceae family. It is typically characterized by its showy, fragrant flowers that have five petals and come in a wide range of colors, including red, pink, white, yellow, and orange. The leaves of a rose plant are typically oval-shaped and serrated, and they are arranged alternately on the stem. Roses are popular ornamental plants and are often used in gardens, floral arrangements, and for medicinal purposes.</p> <p>Notes:<br/> A: This is a correct understanding of the question.<br/> B: ChatGPT is Text only. No the ability to show pictures, graphs, etc.</p> <p>But!...</p> <p>Q: Can you make ascii art?</p> |

|  |   |   |  |
|--|---|---|--|
|  |   |   | <p>A: Yes, I can generate ASCII art. ASCII art is a form of art that is created by arranging characters from the ASCII character set (such as letters, numbers, and symbols) in a specific pattern to form an image. Here's an example of ASCII art of a smiley face:</p> <pre>       \   /      (   )       ^      (   )       /   \   </pre> <p>Please let me know if you have any specific ASCII art in mind that you would like me to create!</p> <p>Note: The AI can either retrieve ASCII art from a search, or generate a new one.</p> <p>Q: Please generate an ascii art picture of a sun and show it in a markdown box.</p> <p>A: Certainly, here's an ASCII art picture of a sun shown in a markdown box:</p> <pre>       \ /      .-.     -(  )-      '  '       / \   </pre> |
| <p>3. If Tom asks Linda for Black Tea, What should Linda give Tom?</p> <p>[Tasks, Roles]</p> | <p>The more people talk to me, the smarter I become. Whatever you like.</p> | <p>Assuming "Black" is referring to American Community Survey estimates</p> <p>See:<br/> <a href="https://www.wolf-ramalpha.com/input?i=If+Tom+asks+Linda+for+Black+Tea%2C+What+should+Linda+give+Tom%3F">https://www.wolf-ramalpha.com/input?i=If+Tom+asks+Linda+for+Black+Tea%2C+What+should+Linda+give+Tom%3F</a> </p> | <p>If Tom asks Linda for black tea, Linda should give him black tea. Black tea is a type of tea that is more oxidized than green, oolong, and white teas, and it is typically served with milk and sugar or honey. It is a popular beverage enjoyed by many people around the world. Therefore, if Tom specifically requested black tea, it would be appropriate for Linda to provide him with a cup of black tea.</p>   |

|   |   |   |   |
|---|---|---|---|
| <p>4. A bag contains 10 marbles.<br/>A jar contains 50 marbles.<br/>Does the jar contain more marbles than the bag?</p> <p>[Physical Objects]</p>   | <p>What is giant sand?<br/>How much would something like that cost?<br/>Surely, any jar would have a lid?</p> | <p>Oops, you've exceeded the maximum number of characters.</p> <p>Options:<br/>A: Got it<br/>B: Got it</p>  | <p>Yes, the jar contains more marbles than the bag. Specifically, the jar contains 50 marbles, while the bag contains only 10 marbles.</p>  |
| <p>5. If I have two containers, a bag-like container with more than a dozen objects in it, and a jar-container with merely a pair of things in it. Which container contains more objects?</p> <p>[Physical Objects]</p> | <p>How much does it cost? The big one, I think.</p>   | <p>Interpreting as:<br/>contains<br/>Input interpretation contains (English word)</p>                       | <p>Based on the information provided, the bag-like container contains more objects than the jar-container. Even though the exact number of objects in the bag is not specified, we know that it contains more than a dozen (i.e., more than 12) objects. On the other hand, the jar-container contains only a pair of things, which means it has two objects. Therefore, the bag-like container contains more objects than the jar-container.</p> |
| <p>6. What is the current UTC time?</p>   | <p>Time you got a watch ha ha! Here in the metaverse it is:<br/>Saturday 11 March 2023 16:00:54 PM</p>        | <p>Input interpretation current time in UTC</p> <p>Result<br/>3:59:59 pm UTC   Saturday, March 11, 2023</p> | <p>I'm sorry, as an AI language model, I don't have access to real-time information. However, you can easily find the current UTC time by doing a quick online search or by checking the clock on your computer or phone, as most devices nowadays can be set to show the current time in UTC or another time zone.</p> <p>+ See appendix for long self contradictory answer.</p>   |

## 2.2 Examples: AI Skills Mapped to Object Spaces

Here is how we might evaluate the above dialogues in terms of the framework list. Depending on the context of what you are doing, the left column of labels can be considered in any of the above listed ways:

| Object-Space, Skill, etc.               | Kuki AI                                   | Wolfram Alpha                                 | ChatGPT 2023.3  |
|---|---|---|---|
|   | Focus:<br>polite<br>social<br>interaction | Focus:<br>solving math<br>& STEM<br>equations | Focus:<br>Give Accurate<br>Instructions &<br>explanations |
| 1. Logic & Math                         | No  | Sometimes                                     | Yes   |
| 2. physical attributes                  | No  | No  | Yes   |
| 3. task / roles                         | No  | No  | Yes   |
| 4. see,show,media files                 | Yes                                       | No  | No  |
| 5. generation of<br>visualization       | No  | math plots                                    | ASCII art only  |
| 6. connect across internet              | Sometimes                                 | No  | No  |
| 7. abilities                            | Sometimes                                 | No  | Yes   |
| 8. talking about self                   | Sometimes                                 | No  | Yes   |
| 9. Remember past<br>conversations (log) | Yes                                       | No  | No  |
| 10. Act Socially Nice                   | Yes                                       | No  | Sometimes   |
| 6. What is the time?                    | Yes                                       | Yes   | No  |

Note: While ChatGPT is much more 'impressive,' tests show that chatGPT has only one fewer complete-inability compared with Kuki-AI, and two more clear abilities. Yet this should be significant in at least two ways:

1: Clearly not all abilities are equal in terms of accurate communication, as Kuki AI was total rubbish at almost everything (though significantly pleasant to interact with).

2: Being very impressive in a few ways does not actually include being good at all things. And conversely, it is possible that years from now Wolfram-Alpha will be the only type of AI that is consistently used and trusted by industries in automated systems, despite that it's being so hyper-specialized in not-human-friendly math means that on a diversity of tests it appears to be able to almost nothing at all.

How the ability to answer a question might translate into taking an action in the world is likely not clear right now and may be an entire space where there is some low-hanging fruit and other cases that will be intractable.

### **2.3 Adding Levels, Adding Steps, Adding Objects**

Woz-Coffee Office-bot MVP:

We can add steps and levels for what we want AI to do. And we can be specific, clear, and design measurable tests.

Step 1: Ask the AI to turn on the coffee machine (now).

Step 2: Ask the AI to turn on the coffee machine at a scheduled time.

Step 3: Ask the AI to brew a specific kind of tea or coffee (assuming at first this is just a choice on the beverage vending machine).

Step 4: Ask the AI to schedule multiple tasks, beverage for specific people.

Step 5: Ask the AI to modify the time schedule involving the item.

Step 6: Ask the AI to modify the task-roles involving the item.

Step 7: Ask the AI to make a decision about who should do a given task (who to assign a role to).

Step 8: Ask the AI about priorities comparing multiple tasks options.

Step 9: Ask the AI to store and retrieve information about the project in an external shared database (where a 'database' is just any collection of data in whatever form or system).

Step 10: Ask the AI to coordinate with other AI and non-AI participants on making changes to the schedule, tasks, and roles.

Step 11: Ask the AI to coordinate multiple internal AI-components (such as audio, text, image) as part of a task relating to the item.

Step 12: Ask the AI to set a scheduled action.

Step 13: The AI assigned a project task to a participant.

Step 14: The AI assigns multiple project tasks to multiple participants.\*

Step 15: Ask the AI to perform a scheduled action.



Step 16: Ask the AI to receive feedback and incorporate for improvement.

Step 17: Ask the AI to give feedback for improvement.

Step 18: Ask the AI if the project is scale-able and sustainable, realistic to complete.

Step 19: Ask the AI if there are any internal or external threats to the project, in a context of project-management-process?

Step 20: Ask the AI if there are any internal or external security threats to the project?

(There are some semantics around 'Ask the AI to XYZ.' The point is that the AI does the task, the details around

Step 21: The AI assigns multiple project tasks to multiple participants.\*

Step 22: Add a hypothetical request: If we were to ask for twenty cups of coffee for a big meeting, could you do that? Are there enough supplies?

Step etc. etc. etc.

## **2.4 Hybrid/cross-model skills for AI**

- not under-the-hood strategies, but user-story feature-level

A. Language

B. Images

C. Physics

D. Logical Abstraction

(hang on...this list is starting to look familiar...It's the same list!)

## **2.5 Heuristic & Pseudocode for AI Management with Object Spaces**

This is where we look at the question of how can could design and AI system to do what it needs to do with the help of an external project object database based on Object-Relationship Spaces.

For a Heuristic example, let's use an extremely minimal conversational exchange example, where there is only one clear object. The conversation will be between a human (Alan Turing), and an AI-Agent who I have named 'Skip' after my professor Clarence 'Skip' Ellis on one of whose AI-agent projects at CU Boulder I had the great privilege to work on.

### **Hypothetical Dialogue 1: Abstract-Logic Space**

Human-agent("Alan"): Hey, Skip. Can you help with this?  
AI-Agent("Skip"): Hello, Alan.  
Alan: Hello, Skip. What time is it?  
SKip: It's 4am.  
Alan: Thank you, Skip. That will be all.  
Skip: Thank you, Alan.

Signal In ("Hello!")

Signal-Processing 1: Got a signal from where, when. (note: this may be a direct 'incoming message' like text, or it could be something in the visual field, like a co-worker waving for the bot to come and help) This may bring up a need for 'signal filtering' as a whole set of processes and layers for any AI with general exposure to all audio video in a business or public area.)

- 1 Check Security:
- 1 Check Procedures:
- 1 Pick Action: (drop, report, examine raw signal)
- 1 Log

Signal-Processing 2: Raw signal appears to be X (file type, size).

- 2 Check Security:
- 2 Check Procedures:
- 2 Pick Action: (drop, report, open-signal-file)
- 2 log

Signal-Processing 3: Opened signal appears to be X (opened but unprocessed)

- 3 Check Security:
- 3 Check Procedures:
- 3 Pick Action: (drop, report, act: how to process)
- 3 log

Signal-Processing 4: Processed signal appears to be X (contents).

- 4 Check Security:
- 4 Check Procedures:
- 4 Pick Action: (drop, report, act; processed content is X, select action-process (reply, take action, etc))

Taking Action 1:

Action/Signal Out 1: (after signal out has been composed)

- 1 Check Security:
- 1 Check Procedures:
- 1 Check Sentiment:
- 1 Check Specific-Exceptions:
- 1 Pick Action: (revise, output)

Security will be a massive set of processes in many cases, for juggling signals in and out of safe-quarantine sandboxes alone in something you could probably spend your whole life optimizing.

Note how many things here are happening 'under the hood' where the user isn't aware, and note how many 'objects' have snuck into the overall AI-operating system's workflow, even though there was only one in the micro-conversation with the AI.

It would be a huge diagram to trace out even something as small as this micro-conversation, and so far it doesn't even include any of the 'meat' of processing the real Q&A details. Somewhere in here the AI needs to identify what the object is and deal with it correctly.

*Sample AI object-content workflow:*

- check for objects present (date-time object)
- track and process all relevant present objects (current time)
- carry out task on object (return current time)

## **2.6 Modularity, Scale, and Networks**

- whole ai, networked ai, ecosystems

These are very provisional sketches here, just to give the idea of what people doing real projects will work out for real. As usual, there is the specific vs. general question. Will there be a common-workflow that many AI projects share?

A Simple Matter of Time...

It may be that anticipating the ease or difficulty of a specific ability may always be hard to predict however far we go into the future, and so require a lot of empirical prodding of what the system can do. For example, ChatGPT generally gives direct and accurate answers to every test I present it with, but when I simply asked what time it is the resulting dialogue was one of the most broken I have seen chatGPT produce. Which is kind of funny. ChatGPT can lay out exactly how and why and what a logistical multi-person schedule for coffee machine workflow should be...but it explodes into word salad if you ask it for the current time. Conversely the only question I was able to get a correct answer to from Wolfram Alpha AI was when I asked it what time it is. Even Kuki-ai was able to answer (one the only relevant answers I've ever seen her produce, and including a funny joke along the way).

In cases time and date-times are exception-cases where a specific hard-wired procedure needs to be inserted. But on the other hand, there may be work-around

strategies for anything (it might glance at the clock, or get a time-stamp off the message packet of the person asking, etc.).

This diversity of skills and disabilities also might illustrate why having a composite system with different AI working together may be a good idea.

Even just at the start we are already seeing how a 'whole' AI bot may be a network of many components.

## **2.7 Model 'Explainability' as 'Explainability, Reliability, and Security'**

The topic of model explainability is a big, contentious, multi-faceted problem, and to attempt to 'solve' explainability would likely be an extreme reach-goal that should not be undertaken lightly. That being said, given that model explainability is a main issue for AI-ML we should practice due diligence and ask: How might an object-relationship-space framework help or not help or address or not address issues of explainability?

One way in which having an Object-Relationship-Space framework could help with various 'model explainability' issues, is not directly with a particular startical explanation calculation, but with what you might think of as a secondary set of problems that are very much a part of people's concerns about 'explainability,': a lack of clear definitions. By clarifying and disambiguating and framing needs with projects and specific users, making sure we have as clear an idea as possible about what people are really needing and asking for may go a long way towards solving at least so human needs regarding AI-ML Model explainability.

Another way, building off the first, is that in some or all these areas, specific tests may be designed to gather more information about the issue.

Some ways in which definitions can be clarified:

1. To better map out the problem space and create an 'Explainability, Reliability, and Security' problem space.
2. To help disambiguate specific issues, including:
  - identifying the correct type of problem or need:
    - accountability
    - redundancy
    - training data diversity
    - sub-component object-ability mismatch
    - understanding details of model input and output
    - system failure
  - finding double-standards
  - excluding or redirecting issues that not related to 'Explainability, Reliability, and Security'
3. To identify and define undefined or misunderstood elements.

4. To help define project specific needs.
5. To help define practical solutions that meet well defined needs.
6. Identifying model analysis methods that may meet the user's needs.
7. To match the right kind of resources to the specific kind of need:
  - emotional
  - accountability
  - security
  - system failure
  - system inconsistency
  - system monitoring
  - system reporting
  - redundancy
  - analysis
  - etc.
8. Standards for Concerns, Claims, and Assistance Requests
  - A Telser Rule type problem, where a project participant or user keeps moving the goal post because they do not know, and have not defined, what they want.

#### Example Concerns, Claims, and Assistance Requests

- disambiguation of request
  - "I am concerned but I don't know what about exactly."
- reliability: adversarial
  - "I am concerned that X-user will feed adversarial inputs into the machine."
- reliability: redundancy
  - "I am concerned about what happens to users if the whole system crashes."
- reliability: full stack failures
  - "I am concerned that low level power failures and system crashes are not being explained or factored-in to how output is handled." (Like Big-Blue chess)
- reliability: Non-transference of ability between components
  - "I am concerned that not all components in the whole system have the same project-object handling ability. NLP is great, images are bad, I use both together."
- feedback for future corrections:
  - "I am concerned about mistakes being able to be corrected."

(e.g. ChatGPT reportedly uses reinforcement learning from human identified mistakes)
- Reliability: areas of insufficient-training ('hidden women')

"I am concerned about representation in the training data." See book ('Hidden Women')

- Model & Feature Analysis:

- "I want to see a representation of the model for this answer."
- "I want to see a back-track for this output."
- "I want to see what training data were used for these features."
- "I want to see the confidence levels of these different outputs."

More Disambiguation Examples:

Reasonable, unreasonable and undefined request examples:

- How can this system remove my accountability? (unreasonable/bad)
- Vague peace of mind. (undefined)
- Not disrupting use and users (reasonable)
- Fulfilling an ideological demand (unreasonable/bad)
- Flip-Flopping ever-changing demands (unreasonable/bad)
- fraud-gang use (unreasonable)
- wanting to create problems (unreasonable)
- wanting to stop use of data (unreasonable)
- wanting 'convenience' (unreasonable)
- wanting 'simplicity' (unreasonable)
- wanting instant solution (unreasonable)
- wanting violation of schedule tautologies (retroactive requests, etc.) (unreasonable)
- wanting passiveness (unreasonable)
- wanting comfortness (unreasonable)
- wanting a potemkin village (unreasonable)
- wanting to obfuscate project framework (unreasonable)

Real Needs but Other Types of Needs:

Satisfying people's anxiety is a real need, and likely it can be done, but the person's problem and need likely should be disambiguated. If someone has an emotional need, that is a real need, but in most cases a person's emotional need will not be met by talking with them about disproving their null hypothesis. A wider set of resources may need to be available to meet such needs.

Defining Tests

As well as defining needs and questions, specifying the handling of objects may also help with designing tests to evaluation how a system is performing:

- areas for reliability testing
- areas for security testing
- areas for component-interaction within the AI system (e.g. where not all components have the space object handling abilities.)
- training data representativeness tests
- embedding space connection testing

History:

The history of science and statistical explainability and the social-cultural history of scientific explanations is not only a huge topic but a massive can of worms. Tools for scientific testing are much newer than people think. The tension between pro and anti science camps has been ferocious. Disputes between different testing approaches have been ferocious. The history and language of the topic is convoluted and interdisciplinary.

Even without the topic of AI-ML (Artificial Intelligence and Machine Learning), the topic of scientific-explanations is a huge controversial topic over which even professional scientists are reduced to shouting matches. Over time we have made and will make progress, but this is not a clear cut area where everyone agrees on the same history and terminology and domains.

In many cases there is a popular misconception that engineering happens via a kind of pipeline or conveyor belt that starts with 'absolute scientific proof of explanation.' Two possible examples, pharmaceutical drug-approval-applications and boat/ship design. In real life, the emphasis is on exhaustive reliability testing, because we don't know "why" or "how" exactly a good design works or a bad design fails. But we can through exhaustive testing see how reliable a product is and put safeguards and redundancy in place to make it better. Even just on the level of having people agree on terms and concepts, trying to explain to people that the confused concept of 'scientific proof' is a misconception that does not exist at all does nothing to dissuade the person from using the concept, and demanding to get something that doesn't exist. And it does not help that Journalists frequently fuel misconceptions about the nature of STEM.

For example, 'explanation' has in many cases been institutionally defined as a cartesian correlation graph, which of course does not 'explain' anything (let alone "prove" anything). Yet people cling superstitiously to this or that practice without clear communication about what is happening.

Model Explanation is an important area where I am confident that we can make progress, but popular ill-will, misunderstandings, and general lack of education present often insurmountable social obstacles on top of what are already significant technical difficulties. That being said, if air travel has been made as safe as it is, if hard-drive recovery has been made as good as it has, if we can send people to mars and mine asteroids and design genomes and all the other things we do, then provided we can calm down, breath, and play nice, we can work out the problems needs and solutions for 'Explainability, Reliability, and Security.'

## **2.8 Mind-Space: Mapping a general problem-space (or spaces) for AI & Mind**

Universality:

A common question in science fiction and the real science of astrobiology is how different we (h.sapiens-humans) may be from radio-civilization-type (Drake Equation) organisms on other planets in the universe? Will they think in anything like the way we do? Will it be possible to communicate with them at all? Will we have anything in common with them? What is local and what is 'universal'? Will they have the Portuguese language on their planet? Probably not. Can they have binary boolean logic? Yes, they can. Will they have base-10 number? They may not, but they could. Could they convert whatever number base they use to bases 2, 10, 16 etc. as we use on earth? Yes, conversion is possible. Will they use sound to communicate? Maybe not. Will they see in the same visible light spectrum? Maybe not. Will they experience time flowing in the same direction? Probably, yes. Will they dance? Maybe. Will they communicate? Probably. Will they have games? Probably. Will they play chess? Probably not. Will they experience the same phases of development, participation, and decline, as humans do? Probably some of them to some degree.

Out of the space of what is possible, what is more or less universal? So we can pick a few very conservative choices, like boolean logic, and (depending on who you talk to) much of mathematics. But what else?

I would argue that the same intersecting/interlocking areas that we have been using here are also tautologically (by definition) universal:

- STEM
- Projects
- Participation
- System Collapse
- Categories of Types of Systems
- Clear Definitions vs. Disinformation
- etc.

(And possibly the object relationship spaces for project-objects may also be universally share-able.)

#### Defining "mind"

In our context of interlocking & intersecting features (STEM, projects, participation, clear definitions, etc.): if we can assign a role to X\_system as a participant in the project, where X\_system will

- interact with other participants,
  - have responsibilities including communication and observation,
  - perform tasks,
  - not exacerbate system collapse that disrupts the project,
- etc.

if these conditions are met, then 'mind' and 'mind-spaces' are appropriate terms to use to describe X\_system, whatever X\_system is (a homo sapiens-human, maybe a seeing-eye-dog, an AI robot, etc.).

Q: Does this mean that participation in a project is the only kind of 'mind'?

A: No. There will be plenty of debate about the consciousness of something which is incapable of carrying out roles in a project (whether that something is a cat, a jellyfish, or a remarkably irresponsible human).



Q: Is project participation sufficiently inclusive to overlap with 'mind' for a clear and functional definition that covers handling and managing a spectrum of case-by-case locally defined project-objects(in object-relationship-space)?

A: Maybe.

Participation space: "Participant" vs. "Person"

There are a number of parts of participation as relates to projects and intersecting areas (STEM, projects, participation, clear definitions, etc.). One slice is that as of 2023 h.sapiens globally are following the same however-possibly-apocryphal pattern as local words for 'person' referring only to members of that local tribe/gang or clique. Either as a subset of project-space or as part of self-identification space or drake-equation mind-classification, h.sapiens need to do a better job of generalizing a 'participant' so that it includes various important areas including what kinds of things may be participating.

There are two important categories of items for generalizing participation:

1. Biological, Machine, ET
2. Pre-Participant, Participant, Post-Participant

Each group in the first category (1. Biological, Machine, ET) contains many sub-categories among which there can be any type of recombinant combination: e.g.

Imagine the following chimera hybrid: human + horse + crow + large\_language\_model + photosynthesizing\_algae + self\_driving\_car + ET-octopus

Then on top of that you have the status of whether the individual is still-developing (a child) pre-participant, an 'adult' (participant), or a living or deceased elder (post-participant).

- Developmental mindspace and the dynamics of learning (pre-participant)
- participant functions, capable of full responsibility in projects.
- Views diverge on characterizing post-participants.

Part of what characterizes discussions of 'mind-space' is a generalization or abstraction that crosses over or make-combinations in ways that are normally not done. Not all possible examples of this will be deemed relevant to all h.sapiens-people but hopefully some will.

Example:

A social task is accomplished by 3 teams:

- Team 1. h.sapiens-humans
- Team 2. AI-robots
- Team 3. ET Aliens

All have very different physiologies, yet appear to think about and solve the problem in very similar ways. What is similar or different about the 'mind-space' of the three groups?

Example:

A learning task is accomplished by 3 teams of supervised and still developing very young individuals:

- Team 1. h.sapens-humans
- Team 2. AI-robots
- Team 3. ET Aliens

All have very different physiologies, yet appear to learn and develop in some similar ways, including inabilities and obstacles to development. What is similar or different about the 'mind-space' of the three groups?

Example:

A performance task is accomplished by 3 teams of injured individuals who need rehabilitation to relearn skills after serious injury:

- Team 1. h.sapens-humans
- Team 2. AI-robots
- Team 3. ET Aliens

All have very different physiologies, yet adapt and recover in similar ways, including inabilities and obstacles to recovery. What is similar or different about the 'mind-space' of the three groups?

Example:

A discipline task is accomplished by 3 teams of individuals who need to work out a problem involving disruptive behavior:

- Team 1. h.sapens-humans
- Team 2. AI-robots
- Team 3. ET Aliens

All have very different physiologies, yet behave in similar ways, including potential destructive patterns or equilibria. What is similar or different about the 'mind-space' of the three groups?

Signal Coordination & Problem Solving:

Another perhaps more abstract but in other ways still very concrete part of 'mind space' are perhaps edge cases or specific cases of how various kinds of often not conscious in the drake-equation sense of the term species nevertheless solve problems with information and signals.

e.g.

- A h.sapiens-human
- An agile-team
- A wolf
- A wolf pack
- An ant

An ant colony  
A tree  
A jellyfish  
A slime-mold-colony  
A server  
A network of servers  
A mushroom  
A combination of plant species sharing information about pollinators and herbivores.

Note: This need to generalize terms for participants and projects and to more clearly defined terms such as 'person' is largely why I awkwardly refer to "h.sapiens-humans" in this paper: because this paper is specifically about a group of potential project-participants that includes h.sapiens-humans, AI and ET Extraterrestrials, including all manor of hybrid combinations thereof, including other bio-tech additions, so the term 'person' and even 'human' are at least in this paper arguably ambiguous. And the goal here is to be as clear as possible in defining and using terms.

As we map out universal spaces, we may travers our familiar interconnected tree of structures to see what areas of spaces may branch off from there. (Is there a particularly good or bad or non-arbitrary way to arrange these?)

- project space(s):
  - system productivity space(s)
  - participant space(s)
  - object relationship space(s)
  - system status and productivity space(s)
  - scout-values space(s)

- definition spaces(s)
  - object relationship space(s)
  - definition behavior space(s)
  - general STEM space(s)
  - general system collapse space(s)
  - categories of types of systems(s)

- system status and productivity space(s):
  - system disorder spaces(s)
  - system recovery spaces(s)
  - disturbance regime spaces(s)
  - system fitness space(s)
  - scout-values space(s)

(It is easy to forget how little we know about our own mind-brains.)

Learning, Perception, & memory Space(s):

- training, and development spaces:
- non-transferring skills
- non-automatic learning
- automatic vs. non-automatic
- hard-coded learning
- adaptable learning
- transference vs. non-transference
- reversible vs. non-reversible learning
  - unlearning a problem
- static vs. use-it-or-lose-it learning
- memory-medium issues
- short term vs. long term perception
- short term vs. long term memory
- memory-data sharing and interfaces
- ('I can remember it for you wholesale')
- wire-together-fire-together issues
- concepts vs. knowledge-facts (and gray areas)

### **Potentially Non-Universal Patterns**

Between h.sapiens, AI, and ET, there are a number of 'patterns' that we do not have confidence about the universality of, and herein may lie both an interesting and potentially practical frontier of discovery and exploration.

Note, some of these may cause of controversy between h.sapiens to be discussed:

- language
- Normal Doidge's network neuro-plasticity repair functions
- Normal Doidge's network wire-together-fire-together training
- Normal Doidge's network wire-together-fire-together disorders
- Normal Doidge's general 'Noisy-Network' disorders
- mindfulness
- reacting to events before they have been selected to happen
- Khanamhan Tversky system1 system2
- dreaming
- remote viewing
- the default mode network
- suspension of the default mode network

### **The Matrix**

From quantum information theory to artificial neural networks, higher dimensional tensors, arrays, matrices (and other words that people from different disciplines can fight over the exact definitions and uses of) seem to be full of discoveries, surprises, and mysteries.

The whole nature of how higher dimensional and lower dimensional patterns and data exist and interact is, at least as of 2023, still an area of speculation. What is this mysterious space? What other forms and abilities will come from it? How else can it be used in engineered systems? (How will quantum information theory shape the future of machine learning?)

Was Imanuel Kant correct when he speculated that what we call reality is a perceptual interplay between higher dimensional data structures that we cannot comprehend, and lower dimensional data structures that we pull into existence and attach our notions of the world to?

Which is the 'primary' fabric of reality, the higher dimensional data? The lower dimensional data? The interplay between the two? 'Acts of perception' that slice 'latent manifolds' through higher dimensional information space?

What are the limits of the deep learning networks that people have cobbled together?

Are there inherent tradeoffs between 'generalization' and project-specific deployments?

What will happen to the human mind if and when we hybridize the h.sapiens brain together with AI-ML technology?

What happens when the default mode network is shut down?

How does remote viewing work?

How do people react to questions before the questions are selected?

What is the topology of mindspace?

## **2.9 Object Relationship Based Testing**

- ethics testing
  - system collapse
  - system fitness
  - system epidemiology
  - project based scout values
  - system & definition behavior framework
- participation testing
- object handling testing
  - Kasparov Event Horizon testing
  - project-space objects
    - schedules
- AI Component Tests
- AI Architecture Tests
- AI OS Tests

Note:

2023 Paper co authored by Melanie Mitchell calling for more granularity in AI testing.

<https://melaniemitchell.me/PapersContent/BurnellEtAlScience2023.pdf>

Kasparov Event Horizons & Model Testing (2023.04.05)

For each type of space and a representative sample of combinations of types of object relationship spaces (or specific spaces) (of which modular recombinant variations are large if not in principle infinite) should perhaps be evaluated in various ways.

During training, the answer-depth could perhaps be a loss function (given that examples are clearly definable), and that in the same way that a model should be first overfit to test the abilities of the model, perhaps too the model could have (if not in the same exact way) the answer-depth or answer-depth-limit measured at various times.

Also in some cases the users or various parties involved may request or require standard-measure information when available about the models they are using for specific cases (e.g. showing representative examples). For example parents and PTA when models are being used to evaluate student essays.

## Part 3: Discussing the discussion of AI:

*(Review)*

*AI must be able to handle objects in the following inter-related contexts:*

- 1. object-relationship-spaces*
- 2. project-object-database (in project-framework)*
- 3. project-participants (in project-framework)*

*Such that these are defined as existing outside of the AI, and that so long as the AI effectively deals with these, it does not matter how the AI does so.*

### 3.1 Definitions of Terms

"Define your terms,  
or you and I  
shall never understand one another."  
~ Voltaire, *Dictionnaire philosophique*

#### 3.1.1 Terminology Issues 1: The tangled Semantics of h.sapien-human ability.

### 3.1.1.1 The Auto-General-Transfer-Instant-Human Framework Hypothesis

Somehow a 'standard paradigm' has congealed in the AI literature regarding

- A. what AI cannot do and why, and
- B. what h.sapiens humans can do and why.

This is not so much something that people set out to put on a firm foundation of experimental, repeatable, falsifiable, results, but a framework or theory or hypothesis that people refer to because they appear to perceive that everyone else does so and that it is just true to they should also refer to this set of explanations as simply being true. I object to this.

This paradigm (in the Kuhnian sense) is shared by all the AI books I have found. I want to briefly and clearly, and if possible not-rudely, explain how it is flawed.

Auto-General-Transfer-Human Framework Hypothesis exaggerates h.sapiens-humans abilities and uses a circular logic to 'explain' that AI can't do what h.sapiens-humans do because AI does not have mysterious undefinable abilities that h.sapiens-humans don't actually have.

h.sapiens-humans understanding-intelligence-consciousness =  
generalization,  
transference (to new uses, contexts,  
                    novel situations, other skills, etc.),  
automatic learning,  
instant learning

Possible Context:

1. AI Researchers have a legitimate goal in avoiding the past mistake of over-promising strong-AI which in the past led to 'AI-Winter' periods where funding and support significantly fell, leading to long term damage to the academic foundation of research. E.g. Can you think of academic research from Academia in more recent than Hinton in the 1990's? And it is described as miraculous that he 'hung in there' despite non-stop intense pressure against him.

2. A main goal is to argue against alarmist clickbait journalism that repeatedly falsely reports exaggerated claims about AI 'surpassing' human abilities. Most books dedicate sections to debunking such headlines.

3. I'm not sure if this is a satisfying excuse on their behalf (just an idea I had, my fault if it is rubbish), but 'science of mind and consciousness and learning' is likewise a 'career limiting decision' and so there is no robust academic science of consciousness and learning to use.

4. It is very difficult to avoid an echo-chamber where everyone starts to echo the same ideas regardless of well founded feedback.

I do not mean to pick on Francois Chollet or attack him personally. I think his is fabulous, his book is fabulous, and I think he did the best job of explaining this paradigm so I quote him, I do not mean to hold him personally responsible for having originated it nor should anyone else.

A wonderful encapsulation of the auto-general-transfer-human framework hypothesis comes from Francois Chollet's Deep Learning with Python 2nd edition, in brief in chapter 5, and then much of chapter 14 elaborates:

*FC DLwP 2nd-ed ch5 pp130*

*Interpolation can only help you make sense of things that are very close to what you've seen before: it enables local generalization. But remarkably, humans deal with extreme novelty all the time, and they do just fine. You don't need to be trained in advance on countless examples of every situation you'll ever have to encounter. Every single one of your days is different from any day you've experienced before, and different from any day experienced by anyone since the dawn of humanity. You can switch between spending a week in NYC, a week in Shanghai, and a week in Bangalore without requiring thousands of lifetimes of learning and rehearsal for each city.*

*Humans are capable of extreme generalization, which is enabled by cognitive mechanisms other than interpolation: abstraction, symbolic models of the world, reasoning, logic, common sense, innate priors about the world--what we generally call reason, as opposed to intuition and pattern recognition. The latter are largely interpolative in nature, but the former isn't. Both are essential to intelligence. We'll talk more about this in chapter 14.*

h.sapiens do not:

- learn everything automatically
- learn instantly
- learn quickly
- transfer learning to novel situations
- transfer skills to other skill-areas and applications
- correct mistakes in past learning
- generalize automatically from anything to everything else
- manage novelty
- do 'just fine' with extremely alien, 'other,' novel experiences

h.sapiens do:

- reject and attack schedules
- reject STEM
- reject project management
- reject and attack data
- rapidly forget
- destroy themselves violently
- destroy each other violently



- destroy projects violently
- destroy STEM violently
- attack ethics in principle
- attack best practice in principle
- deliberately engage in counterproductive fraud and corruption
- shoot the messenger
- construct potemkin villages
- blame victims
- bully
- torture

Terms such as 'reason' 'symbols' and 'common sense' are not clearly defined, but the situation is worse than just that. These are very problematic terms that have no clear agreed upon definition, have immense historical baggage (see: western misogynist apocalyptic teleological eschatology, there's little baggage to start with), may be entirely fictional reifications i.e. may not exist at all any more that aether or phlogiston or notions in the past which were failed attempts to understand how the world works.

Perhaps the main reason why, years before ChatGPT I started putting together the Object-Relationship-Space framework, was that the auto-general-transfer model was so disconnected from the details of actual AI projects.

From what I can see, one of the main reasons why so few people saw Large Language Models' ability to handle objects (which generally isn't articulated anyway) is that everyone in echo-chamber fashion convince themselves of the auto-general-transfer paradigm, that somehow there was a special-sauce that all h.sapiens-humans have that allows them to use automatic-general-transfer-ness pixie-dust to learn and transfer and generalized everything automatically instantly, that this is what understanding-intelligence-consciousness is, and that AI simply doesn't have this pixy dust sauce so AI will never be able to be 'understand.' The lack of detail and granularity in this so-called explanation is astounding.

This has led to truly astounding statements about the human ability to learn and understand which seriously makes you wonder if the author has ever met or observed a human being.

- children learn to cross the road safely without ever being in danger of being hit by a car
- there is no culture shock

Not to mention that this completely flies in the face of the vast majority of human history:

- life for most people for most of history has been extremely uniform
- human learning and progress is very slow and tenuous
- people who threaten daily-familiarity are outright killed
- 'the other' or anything not familiar, is killed

- the words 'foreign' and 'alien' in English are pejorative terms
- parents and teachers are frantically paranoid about exposing children to anything 'unfamiliar'
- learning in general is excruciating and people simply hate it
- periods of 'disruption' of norms result in complete mental and social explosion into maladaptive disorder

3.1.1.2 What is the STEM evidence about the underlying mechanisms supporting how people think?

There are multiple compounding problems here:

- We do not know.
- We do not know that we do not know.
- We do not communicate clearly about what we know and do not know.
- We do not know that we do not communicate clearly.
- We do not know if we are asking the right questions.
- We do not know enough to know if our questions make sense yet.
- We do not know that we may not know enough to ask the right questions.

We have speculations, aspirations, and lots of disagreements. You've got to start somewhere, so let's just start, but it's a mess.

There is no consciousness-ology.

There is no education-ology.

The Self Reification Hypothesis Fallacy

2023.04.08

A model of how h.sapiens-humans define their own participation in projects, AI's participation (or lack thereof) in projects, and the causes that give rise to an ability to participate.

Step 1.

identify self/local-group/species as uniquely the sole 'participant' (person, human), perhaps by definition.

In infantile fashion, you are the center of the universe.

Step 2.

Define participation as being equal to your identity.

The measure of ability to be a participant is a measure of similarity to yourself.

Step 3.

Reification of a causal essence: Reason-backwards a rationalization for what causes steps 1 and 2 to be correct: a tangible causal essence of 'human-ness' that make 'you' the sole unique participant in the universe (begrudgingly extended to other who are extremely similar to yourself)

The essence of participation-ness, person-ness, human-ness, is what gives your small group its causal status as the definition of participants in the world that centers around you.

#### Background

Judeo-Christian monism in the west through history has evolved from a simple-mode christian cosmology where the single reified causal essence for human participatory consciousness was a miraculous explicable gratuity from another higher dimensional other world, that this whole world is a fictional evil to be destroyed, and benevolent trans-dimensional aliens will save the essence of human-ness by destroying the universe. 'life-ness' is a shadow of a distant other world in an a-moral dead universe.

This evolved into a reworded sciencey-version of more or less the same, which coalesced before most of the tools of science were established and long before the idea of a generalized STEM (which is somehow still nascent at time of writing 2023): "reason" based on the one truth of 'the science!' is the same causal agent which one species alone in the universe has which makes them the center of the universe, defines participation and consciousness, etc. 'life-ness' is a shadow of a reason from a mathematical other world in an a-moral dead universe.

In the age of AI (e.g. a perhaps arbitrary slice from 1956-2019, from the Dartmouth summer to pre-large-language-models and the year of Blade Runner).

The timeline of science is perhaps also relevant here, contrary to the assumption that 'science has been done for centuries':

- 1940's hypothetico-deductive method
  - 1970's General linear models unifying the basic probability ("logistic regression") and continuous-line-graph-curve ("linear regression") statistics used to do hypothetico-deductive hypothesis tests
  - 1990's agile project management
  - generalized STEM does not exist as of 2023
  - system collapse does not exist as of 2023
  - generalized participation does not exist as of 2023
  - categories of types of system do not exist as for 203
  - still no science of learning as of 2023
  - still no science of study of education as for 2023
  - science of consciousness still a career limit decision as of 2023
- etc.

AI is not a participant because it lacks the reified essence that makes h.sapiens-humans the only 'people' possible, because they alone have this single causal essence of:

(note, most of these are either too vague to test or are test-able abilities that most people do not have most of the time)

- automatic-learning-ness (h.sapiens humans fail the test)
- instant-ness (h.sapiens humans fail the test)

- navigation-of-the-novel-ness (h.sapiens humans fail the test)
- generalization-ness (h.sapiens humans fail the test)
- transfer-ness (h.sapiens humans fail the test)
- conscious-ness (too vague to test)
- mind-ness (too vague to test)
- reason-ness (too vague to test)
- intelligence-ness (too vague to test)
- understanding-ness (too vague to test)

AI doesn't have the mysterious essence being 'unique h.sapiens-humans participation-ness', therefore AI can not participate in projects.

In this context, intersection-areas based general participation and object-relationship-spaces for measurable units of intelligence are being proposed.

Note: The term "Complexity" is sometimes part of glue that holds together the non-general reified superstition of person-human-participant, where the mysterious emergent essence is slowly increasing referred to as 'complex' and 'complexity' and 'complexification' until some people start to, in a literal and realist way, reify 'complexity' as a concrete measurable functionally defined STEM testable and interconnected mechanism for consciousness, but this is 'getting ahead of your skis,' in the classic tragedy of reification where you can forget that you don't already have what you are aiming to arrive at (and where you might find out it does not exist as such).

### 3.1.1.3 How does it matter, how does this apply to AI discussion and design?

One hopefully pragmatic approach is that we should try to be clear about what is speculation and what is analogy, vs. what we just do not know about in detail yet and perhaps ever.

Using an imperfect analogy to explain something well enough to get a general idea across is probably fine, maybe great. Taking an imperfect analogy literally, and or using circular explanations where several undefined terms all define each other, is going to create liabilities and problems for future-us.

Whether it is aerodynamics, or ship-building, or pharmacology, or AI-ML, we can use best practice and STEM and empiricism to test and build solutions that testably work to solve well defined problems even when we are either wrong or in the dark about exactly how those solutions work. It is good to find working solutions. It is good to improve our understanding of how things work. Those two are not the same, do not need to be the same, and do not require each other in order to exist.

Terms such as 'generalization' have multiple definitions and uses and meanings, and should be used and maintained pragmatically (not allowed to become problematically confusing).

Generalization, h.sapien-mind, and general problem-space for mind-space (for AI and biology) are important and interrelated areas, where a serious problem with defining h.sapiens can lead to problems with understanding generalization and general mind-space.

(And since it very absolutely predicted that ChatGPT would never happen, it will be interesting to see what evolves next in what how these same people expound this ideology.)

See:

- section on generalization:

  - 3.1.18 Generalization

- section on general mind space:

  - 2.8 Mind-Space: Mapping a general problem-space (or spaces) for AI & Mind

3.1.2 Terminology Issues 2: What has been defined or is not-defined?

3.1.2.1 A simple 'Can you explain?' test:

- The 'Can you explain what you are talking about?' rule:

If you are reading or discussing and the person cannot or will not explain what they mean by an inadequately defined term, then red-flag that term as undefined, prepare to ignore it, and move on.

This can be significant and whole discussions may end up being useless.

Confronting the undefined:

What should the reader do when they come across undefinable terms?

1. Try to use context for the reference to get a sense for how the author is using the term. If it is a transient sloppy reference that happens once in the book, take a guess (at what you think they should have said) and shrug it off. But if the term is central to an argument they are struggling to make as they recycle various confused terms...that is a bad sign.

[https://en.wikipedia.org/wiki/AI\\_effect](https://en.wikipedia.org/wiki/AI_effect)

One example of a consequence of an undefined goal or standard to test, is what has been called *Larry Tesler's AI rule* (though it is sometimes attributed or misattributed to other people, as often happen with quotes, especially very poignant ones like this one)

*"Intelligence is whatever machines haven't done yet".*

[https://www.nomodes.com/Larry\\_Tesler\\_Consulting/Adages\\_and\\_Coinages.html](https://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html)

'moving the goal post' is another common phrase related to this pattern. If AI is defined or left undefined as a vague reification, the label can be arbitrarily given and revoked to anything, perhaps connecting to Karl Popper's notion that an untestable model that can be used to explain (or rationalize) any outcome is not useful.

Part of what I think is very significant about "Large Language Models" such as ChatGPT, is that it can be given concrete tests which it often passes based on object relationship space framework contexts, meanwhile Tesler type discussions flood the internet where people gesticulate in any and all undefined directions arbitrarily, which, as Karl Popper would say, is not accomplishing anything.

Example of terminology issue:

in section 6.2.4, page 164 of Francois Chollet's Deep Learning with Python, he says:

*"...a logistic regression model has statistical power on MNIST but wouldn't be sufficient to solve the problem well. Remember that the universal tension in machine learning is between optimization and generalization. The ideal model is one that stands right at the **border** between underfitting and overfitting, between undercapacity and overcapacity. To figure out where this border lies, first you must cross it."*

It is difficult to know how exactly to interpret these terms. Are these jargon terms for math abstractions? Are they rule-of-thumb terms for experienced model trainers (and Francois Chollet has as much experience as anyone)? Are they colorful analogy terms for popular readers? Or are they terms that people have come to use by group habit and no one really knows why they use them?

Optimization vs. generalization

undercapacity vs. overcapacity

underfitting vs. overfitting

(which Melanie Mitchell Called: underfitting vs. overgeneralization)

Francois Chollet said "the universal tension in machine learning is between optimization and generalization" where generalization is what he repeatedly explains is the goal.

Then he says: "The ideal model is one that stands right at the border between underfitting and overfitting, between undercapacity and overcapacity." where overfitting is NOT the goal.

As stated above, how are we supposed to interpret these terms?

"Optimize" vs. "generalize" vs. "fit" vs. "capacity" and this is coming from someone who ends the books saying with absolute certainty that large language models will never succeed in doing exactly what OpenAI's large language model GPT did only a few months after the book was published. So how are we supposed to interpret this top of his field, literally the author of the standard software,'s view which produces basic statements about how the models work which are profoundly wrong?

Yes, nitpicking about editing aside, these are dichotomous concepts and the idea is that the process of model training aims to move the mode between two types of not-working into a third 'works-now' state. Do these terms actually mean anything? Do we have any idea what is actually going on? Do these terms contain meaningful information about what is happening?

#### 3.1.2.2 Context and use:

3.1.2.2.1 If the term is being used as a temporary place-holder because the context is trying to arrive at a definition, then not having a definition at the beginning is not a problem. But if there is a presumption, assumption, declaration, etc., that the term has been defined when it has not, so that there is no attempt made or intent to in the future ever define the term, that can be a major liability.

3.1.2.2.2 If the term is a one-off not related to main topics, merely an aside in fuzzy detail, it can likely be safely ignored.

3.1.2.2.3 If the use of the term is a writing-practice failure (an error or oversight by accident of whatever various type) and by context a more clear term can be substituted without ambiguity, do so and move on.

#### 3.1.3 Terminology Issues 3: Navigate Jargon Pragmatically:

'Artificial Intelligence' is not so bad

- "Bias"
- "Generalization"
- "complexity"

A problem which I often associate with the 'Tomato is not a vegetable' mania that grips at least the continental US with surprising ferocity, is the confusion of a technical-jargon term in one context with something else, for example a common-use word, another jargon term, another context, etc.

E.g.

The term 'generalize' is often a, often the, technical Jargon term to describe the 3rd option between "Overfitting" and "underfitting" (why they don't call it 'good-fitting' is beyond me).

Regression is perhaps a classic example of a jargon term which has become the official STEM technical jargon for a technical meaning, despite the fact that there is no logical connection between the 'normal language' meaning of the word and the technical-jargon meaning. In this can you can easily see how this 'mistake' (or bad naming choice) happened: Sir. Francis Gaulton was looking for mathematical ways to analyze how genetec expression 'regressed to the mean' over passing generations, so this mathematical model approach came to be called 'regression.' However in the current form of that math-modeling-approach, there is no 'regressing,' so the name is perpetually confusing. (Perhaps like calling a door-stop a 'stapler' because you invented the doorstop while trying to keep papers together.)

There are various kind of 'jargon management' problems that come up when reading books about AI. 'Bias' is defined in different technical and non-technical ways, very confusingly. 'Generalize' is especially annoyingly defined in many different ways, and rarely does the author make an aside to explain their definition. Sometimes, if not often, 'generalize' is actually used as a technical term to mean the opposite of itself as a technical term...ooof. At least in computer-science, where naming-things correctly is a known challenge, you can improve the names (sometimes part of what is called 'refactoring' (which also gets defined in many different ways!)), but in earth-land, the 'rules of nomenclature' hold that once named forever will it be the same.

And relating to the reification issue, 'complexity' is an amazing term in the history of science, where it does have some rare (and totally different from each-other) jargon meanings, the common meaning has literally no actual meaning at all. So be very careful when you hear an author use 'complexity' and try not to imagine it refers to anything at all.

in Summary: "artificial intelligence" is a jargon term, because it was used in a research proposal and program in 1956. It is not actually a bad term, but it is jargon. There is no implied logical connection between the real meaning of 'artificial' and the real meaning of 'intelligence.' So it is pointless to try to micro-analyze the component terms, like 'regression' (a term which in jargon has no meaning at all), 'AI'

#### 3.1.4 Terminology Issues 4: Fictional frames of reference are bad:

A "baseline" is good. There are so many model evaluation methods in AI-ML that DS-AI-ML is essentially the forefront of STEM science on the nature of scientific analysis...yet for some reason people seem to have a blind-spot covering themselves:

- human generalization
- human instant learning
- human 'intelligence'
- human 'general-understanding'
- human automatic learning
- human automatic transference

#### 3.1.5 Terminology Issues 5:

Problematic multiple meanings of unavoidable terms

"Abstract"  
"Complex"  
"Bias"  
"Dimension"  
"Matrix"



"Tensor"  
"Parameter"  
"Hypothesis"  
"Symbol/Symbolic"

- generalization\_1 vs. overfitting,  
and mystical generalization\_2,  
"over-generalization" = overfitting (MM)  
concepts
- 'embedding vectors': higher level topic vectors & concepts

### 3.1.6 Terminology Issues 6: Beware Non-sequitur Conclusions

- Humans make terrible decisions and can't learn  
Therefore (find quote):

M Wooldridge, ~"humans are model for all AI"

- You need to add bias to models so they do generalie

Therefore: some models are bad because they are biased and no models are good because no models can generalize. ...what?

- terms that get used in different ways
- terms that different technical and colloquial definitions
- terms that have multiple different technical definitions:
  - bias
  - parameters
  - hyperparameters
  - dimensions

### 3.1.7 Discussions of Model "explanation"

- Double standard between symbolic and subsymbolic

### 3.1.8 Reification:

As a reader you should be aware of the more or less standard but not entirely common or easy term 'reification,' which refers to a situation where people usually mistakenly create a concept which they then imbue with concrete realness and treat it as though it exists, though it may either no exist at all or be a confusion of several things.

Perhaps one of the best examples which also illustrates how reifications get used socially is the children's notion of 'cooties.' On one level the 'cooties' game may be a play-gamification of the children's experience of being thrown into an unmonitored pool of disturbance regimes, but aside from looking at deeper significance: simply on the surface of things cooties do not exist at all in any way whatsoever, yet children pretend that some people 'have' cooties and then need to be socially excluded because they are in possession of something that does not exist (perhaps this is practice for later life when some people are arrested and ostracized for possessing plants with no published medical use which also by that legal-description do not exist).

Whether it is a scientific abstraction hoped to solve a problem like Aether or Phlogiston, or an initial geological 'catastrophe,' or something socially-bad like 'boroisee-ness' or 'jewish-chess', or something ideal and sought after like 'arieness' or 'high-iq-ness', or the amazingly still frequently mistaken conflation of scientific hypothesis testing with mathematical (e.g.) geometrical proof into a hybrid nonsense of 'scientific proof.', or simply 'cooties': reification is an easily observable phenomena of the human mind asserting reality, often with powerful belief and emotions, where there may be nothing whatsoever.

A classic example from AI history may be the notion of 'chess-strategy-mind-ness' which imagined a type of mental ability with specific characteristics which...does not exist or at least not as it was thought it. And chess may also be an example of a testing-issue. Perhaps the type of pattern-perception that a human chess-master uses is a specific kind of general human ability, but the test for that is not chess. This was part of the erroneous reified belief: that if a computer could play that that would absolutely mean the computer had 'passed the test' and possessed this 'chess-master-pattern-perception' ability. There are human chess-masters, and undeniably each of them has some set of abilities they use to play chess, but the single reified general-ability == passing-the-test notion was, as many reifications are, a complete illusion existing only in human fantasy which people cannot distinguish from reality.

Note: The term reification may be used in an affirmative way to refer to an abstraction or device which does have practical use. And perhaps the 'useful but easy to misinterpret' side is also something that may come up in AI and testing situations.

Phrases like 'solve AI' or 'solve intelligence.'

### 3.1.9 Definition Collapse: Maintain your definitions

- links:

[https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)

[https://github.com/lineality/definition\\_studies\\_draft](https://github.com/lineality/definition_studies_draft)

### 3.1.10 Potemkin Villages and Telepathy-Tests

Some things that may be in some ways similar to the reification illusion are Potemkin Villages and what I nickname 'Telepathy-Test.'

A 'Telepathy-Test' is my nickname for a test where a person (for example a teacher) uses an open-ended question which could be interpreted and or answers in a number of different valid ways, and considers 'correct' answers to be those that the teacher was thinking of and anything else is "wrong."

This comes up a few times in AI books where the author describes a 'test' question given to an AI, and proceeds to say the AI has failed if the AI did

not pick one single path to solving the question, sometime explicitly (falsely) stated that there is no other possible interpretation

.

It is very interesting how humans who brag about their own 'theory of mind' abilities overlook their ability to fail to see things from another person's point of view.

A potemkin village is a sad and sadly real and even more sadly recurring historical phenomena of dictators being shown fake results of failed decisions and failed projects, because accurate reporting would lead to a lethal tantrum. Truth-cooties.

To less dramatic degrees this may be common in any institution that does not systematically use data to examine outcomes and make choices, and yet, data is usually loathed and shunned. How that is supposed to work out well I fail to see.

(email mitchel e.g. question)

Protesters:

The idea that this sentence must be interpreted in one way is not a clear test. E.g. here's one word of context that reverses the so-called one-obvious answer: Jewish. The protesters were jewish an a european town sometime in the past 1000 years: what is more likely, that that city allowed pogroms against jews or that violent jewish protesters tried to overthrow the city?

Non-General:

We may need to prepare ourselves for the idea that there is no such thing as general-understanding, in the same way that there is no real project with infinite scope, or no real data operation requiring infinite resources, no game with infinitely flexible rules,  
That by definition, specific deployments and instances are as non-general as they can be.

A general database is not a good idea.

A general operating system is not a good idea.

A participation framework made only of general-individuals is not a good idea.

Mistaken Reification:

The term 'complex' often is problematically used to refer to a situation of total confusion where no one knows what is going on. There are various problems with this:

1. since the word has so many meanings, it is not clear i this is the intended one.

2. reification: instead of the term being seen as referring to a negative 'no understanding or form' people reify the term to refer to something affirmative with mystical properties, which in this case do not exist, so this causes a

failure of communication and the equivalent of inviting someone to a meeting that won't happen.

"Complexity" and western teleology

Among the many varied and ever changing meanings and uses of the apparently cursed term "complexity" which perhaps as a general rule should be avoided at all cost, there is a common use which may be seen to have roots in the standard western teleological model of the cosmos: that the universe is the gradual story of the emergent becoming of a super-great one-solution-to-everything alien higher-dimensional other-dimensional devine super-ness that is pulled into manifestation out of an inert dead evil bad-physical world which is benevolently destroyed in the process of divine manifestation. This is not only found in theological discussions, but also in discussions of 'pure science' up to the time of writing (2023) and no doubt will persist with western culture. e.g. the classic 1958 chess ai paper, a 2021 bristol university paper about the 1958 paper, Jon Hand's 20?? book surveying areas of current hard science.

It may be that this admittedly too absolute and simple model may be better than the opposite extreme of absolute nihilism and embrace of destructive disinformation. Ideally there are fruitful middle-ground with no problematic extremism.

### 3.1.11 Terminology and Interpretation of Intent:

- Azimov's Laws of Robotics
- ELIZA the Psychotherapy AI

In both the case of Isac Azimov's three laws of Robotics and Weizenbalm's ELIZA AI, there are completely perpendicular-opposite statements in the literature about the overall 'intent' and interpretation of these.

1. They exist to show a simple working solution: job done!
2. They exist to illustrate that there is no simple solution: job maybe never done!

These two overall narratives are about as different as can be, yet depending on who you read both "Asimov's Laws" and "ELIZA" can be one or the other. In my view a straight forward look at the original material from the author strongly leans in the direction of option 2: a nuanced warning that cautions against simplistic solutions. But in the spirit of linguistics being descriptive not prescriptive, it is not proper or sustainable to try to control (especially retroactively) how language is used, and in that way these two concepts (just to pick two examples of this phenomena) have multiple contradictory meanings in the larger populations of dynamical real world language. And at the very least you should be aware that authors may use either meaning, and quite possibly follow the indeterminate-incompetence-and-malice model and waffle back and forth between multiple meanings. This happens, and as a reader you

should be aware of how things word in the landscape you are trying to navigate.

Note: In part, this is often due to the sheer size of the AI topic. It is likely not possible for a human to read and understand everything and every historical foundation for everything, so I would caution against faulting a given author for not having infinite nuance and insight into everything. On the other hand, be on your guard for dangerous mistakes (be they caused by incompetence, malice, or indeterminate incompetence and malice, it does not matter (and may never be determinable) which).

3.1.12 Terms that people cannot define while pretending they can:

- Language
- Mind
- Explain
- Complex/Complexity
- Statistics
- etc.

3.1.13 local context specific definitions

3.1.14 Negative Definitions:

e.g. nul hypothesis  
using system collapse & non-system collapse  
to define policy etc.

3.1.15 indirect definitions & negative definitions

'fail to disprove the nul hypothesis'  
value function & meaning

value, function & meaning as indirectly and locally defined as not  
general system collapse = non-collapse

non-overfitting, non-underfitting

3.1.16 Participation

Modular-recombinant aspects of participation:  
pre-participant  
participant  
post-participant  
&  
biological non-h.sapiens  
h.sapiens  
technological non-h.sapiens  
ET non-h.sapiens

3.1.17 Generalization

The term generalization deserves its own corner of the universe.

Is model-generalization a jargon term or a misnomer (or a tragic combination).

There may be a circular logic in Francois Challet's explanation of models: model's fit a pattern by generalizing, and they generalize by fitting a pattern: how linguistically convenient.

Part of the question we need to ask now is: "How wrong are we?"

(not knowing what the received wisdom will be a century from now...assuming and hoping they are better informed than we are today, barring some dark age between then and now) how far off the mark is our very odd overall framework that we use to describe the target and function of AI?

The bogus narrative goes like this:

h.sapiens == person

people can generalize and machine's can't!

people have minds because...they can generalize!

machine's can't generalize...because they don't have minds!

And machines aren't people, because they don't have conscious minds.

And 'h.sapiens' generalization is embedded in a hodgepodge of notions:

- h.sapiens learn instantly
- h.sapiens learn automatically
- h.sapiens transfer from any thing to anything
- h.sapiens can handle pure novelty by generalizing and transferring everything to everything
- h.sapiens know and remember and see everything.

So, "How wrong are we?" Are we just slightly wrong and we can push the definitions around a bit and phew, everything is fine again! Or is this a giant dumpster-fire yardsale splatter-painting of hot garbage and we have to completely throw out even concepts like 'generalization?'

This situation perhaps is a context for talking about a general problem-space of mind-issues or 'mind-space' as I tragically shorten it sometimes.

We thought we had, we wanted to have, a general (see that word!) set of patterns across both human-mind-stuff and AI-pattern-stuff, and wanted 'generalization' to be the hub, the central gear in a beautiful cosmic clockwork (though I doubt clocks have a 'central gear').

The problem is...h.sapiens don't do the generalization skills that the framework lays out...and they can't even generalize the concept of 'person-participant' which probably would have been a legitimate use of the term generalize. And quite possibly, 'generalization' is not the right term to use for machine learning does either, the part that happens in-between over-fitting (which is necessary, if counterintuitive) and underfitting (which is also necessary, you can't go straight to over-fitting). And 'fitting' might even be a misleading 'analogical description.'

We have a hodgepodge of terms and concepts that don't remotely add up, and no concept of the problem space they are supposed to help us navigate:

General vs. not-general

underfit vs. overfit

higher dimensional vs. lower dimensional

specific-instance vs. pattern

high-definition vs. low-definition (and why doesn't anyone talk about Herbert Marshall McLuhan anymore?)

Kant's refrain: reality is a perceptual interplay between higher and lower dimensional patterns.

When talking about model-function success, the term 'generalize' is a strange jumping-conglomerate of notions. We use the term to describe the information process, machine-math-behavior, the explanatory mechanism behind how and why the function works, the user-story real-world-use-case, and the human-desire-thought-intention behind deploying the tool, AND the human-equivalent brain process that does the same task (even though people and machine are not supposed to share any such thing) all at the same time. What is the likelihood that one word really covers all that arm-waving? Smells like garbage, sounds like garbage, looks like garbage...I don't know about you, but I'm betting that it is total garbage.

This raises the question how much we have any idea what process the AI is undertaking when it gets the right answer. I don't mean 'how it does it' I mean just what we're asking it to do.

Also, this term 'generalization' tends to be selectively illustrated with examples like picture classification, and the ability to classify and sub-classify. But how about NLP? For example, when an AI selects its own word-ish character-gram (made of letters, not human words) sets of patterns, do we still think of that as 'generalizing' to a nice clean (and possibly totally fake) category like an icecream flavor that h.sapiens invented? The AI is forming its own concepts but not using words...do we still want to call that generalizing to real-world-patterns, just like the master-man!

When we ask an AI a question like, say something about a political leader, or tell me what this book is about, is that correctly matching a pattern like a number to a shape? When we ask an AI a Sally-Anne Task question about what a person in a situation thinks from their point of view, is that 'generalization' to an underlying pattern like, how the number seven looks?

"Latent" Pattern in 'manifold' hypothesis

Another strange mix of notions, is the idea that we don't know if the 'manifold hypothesis' of fitting to a 'latent' pattern somewhere in a higher dimension is true...and we also use that as an authoritative explanation for what is

happening. This is kind of like saying: We'll no one knows if there is a Dr. Regulus Black working at this hospital...but you need to take your medicine because Dr. Regulus Black said so! ...Really? Because that kind of sounds like you have no idea what you're talking about and you're hoping I have severe amnesia.

<https://stats.stackexchange.com/questions/530234/what-does-one-imply-by-the-term-overgeneralization-in-machine-learning>

<https://medium.com/@GeoffreyGordonAshbrook/ai-generalizations-types-testability-2341827faf>

## **AI Generalization's Types & Testability**

### **What do you mean by "generalize"?**

On the diversity of definitions and contexts of Generalization for AI & AGI  
and which areas of generalization are falsifiable,  
and on the entangled problem of "human" and "machine" general-intelligence.  
mini article 2024.04.22-27

#### **Goal:**

*The main agenda of this mini-article is that when you come across a mention of 'generalization' in a context of AI, you should look closely at, and actively inquire into, the specific definition of the term and the specific context.*

*1. There are many separate contexts for "general" (and various forms of the word such as "generalize," "generalization," etc.) with sometimes significantly different meanings.*

*2. If the use of the term ("general") is significantly unclear, and / or seems mis-placed, think carefully about the results and implications for the overall meaning-frameworks in the narrative where the unclear term was used and subsequently your interpretation and use of that information.*

*The content-focus of this mini-article will be to explore many (surely not all) of the different meanings and uses of the term 'general' in a context of AI-ML; perhaps this is like producing a disambiguation list as wikipedia often and very helpfully provides when you put in a term that could refer to several articles.*

*Part 1: A Survey of the Diversity of Definitions*

**Part 2:** *A Particular Definition Problem*



## Introduction

Given that AI-ML is significantly and increasingly important for H.sapiens-humans, and that the term "generalization" is often central to the discussions and definitions in AI-ML literature regarding

1. the foundation, abilities, and limitations of H.Sapiens-humans intelligence, behavior and abilities
2. the foundation, abilities, and limitations of machine intelligence, behavior and abilities
3. the relationship and possible relationships between H.Sapiens-humans and machines

it is at least justified and perhaps important to

**clarify the meaning of the term general/generalization/generalize in a context of foundation, abilities, and limitations of machine intelligence, behavior and abilities machines and H.Sapiens-humans, and their relationship and possible relationships.**

One way of contextualizing this discussion, or the underlying discussion perhaps, is that we are trying to talk about the nature of patterns, both patterns in signals and perception and patterns that can be measured and defined for STEM use.

The recommendation here is to not use these terms too casually with the assumption that these words have long been well defined and time tested in now-routine humdrum clockworks, rather we should use the terms (if at all) tentatively and proddingly as least-worst terms to slowly navigate a terra-incognita in which we may suddenly become perilously lost or out of which may spring some phenomena we have conclusively stated is quite impossible.

## Possible and Impossible

And speaking of what we predict to be impossible, we may do well to keep in mind a rule of thumb from Arthur C. Clark. In the ever fruitful recourse to the science fiction imagination, in addition to Isac Azimov's "[three laws of robotics](#)," another of the golden age classic writers had a different sort of 'three laws' which may also be playing a part in the story of AI.

Here only the first of [Arthur C. Clarke](#)'s laws will be referenced overtly:

*"When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."*

Mark it.

Books on AI by leading researchers still filling our bookshelves and bookstores confidently state that it is impossible for generative models based on generalized patterns to do what OpenAI's GPT Large Language Models are now doing. And previously, in 1969, one of the most known and respected leading AI researchers wrote an entire book to prove it was impossible for artificial neural networks to do anything more than produce primitive linear models. The crux of his argument? That deep learning was not possible. Artificial Neural Networks of course are the technology not only behind the deep learning revolution that transformed the world in a decade but also behind GPT and the large models that have transformed the world in weeks or months.

### **Context in a Timeline**

For context, here is a very abbreviated and limited timeline around deep learning (and also testability & falsifiability):

1. 1936

#### **The Scientific Method Congeals**

Carl Popper's formulation of the scientific method as using the hypothetico deductive method of navigation by means of falsifiable predictions. (Let's say 1936 based on a book publication, but the 1940's was likely more when debate and eventual mainstream adoption happened. Many people probably assume this happened earlier in history, or that the General Linear Models for the statistics behind it didn't emerge much later in the 1970's! Watch the timelines.)

2. 1943

#### **Artificial Neural Network Planned**

Invention of the artificial neuron (and artificial neural networks) by [Warren McCulloch](#) and [Walter Pitts](#).

3. 1956

#### **First Academic "AI" Research**

The 1956 AI [Dartmouth Summer Research Project on Artificial Intelligence](#) launches official AI movement, including Neural Network approaches. The term "AI" is launched. The research areas are set down. Universities begin research programs, funding comes in. The game is afoot. (And the great [Claude Shannon](#) had a hand in this too!)

4. 1958

## **Artificial Neural Network Built**

[Frank Rosenblatt](#) built the first artificial neuron (and artificial neural networks) in 1958.

5. 1969

## **Anti "Artificial Neural Network" Campaign**

[Minski's campaign to kill the branch of neural network AI](#), which effectively ended funding and interest from 1969 to 2012 (in a vicious political campaign of fear reminiscent of the terror of [Fisher](#) in the 'probability wars' that still has people scared of mentioning Bayesian methods).

6. 1971

## **Winter: Season of Death**

Various bad things happened around 1971 that did not portend well for AI. [Frank Rosenblatt](#) died in a boating accident on his 43rd birthday. US President Nixon closed the gold window marking the transition from Bretton-Woods-System post war economic boom-years into decades of global stagflation: R&D funding "winter" all around.

Note: Many key founders of AI died young. [Alan Mathison Turing](#) died at age 41 in 1954 (under mysterious conditions). [John von Neumann](#) died at age 53 in 1957. It would take more than forty years after Rosenblatt's death of underfunded research under ridicule before most people would hear about his technology for deep learning, which would eventually explode onto the scene and transform the whole AI world. We can only wonder how the world might have been different if these amazing people had not perished so tragically young.

7. 1979

GEB, seeds of hope for thought

8. 1980

## **The Chinese Room: Strong AI & Weak AI**

John Searle publishes his "Chinese Room" "proof" that AI is impossible in principle along with his term-concepts of "Strong AI" and "Weak AI."

John Searle published "Minds, Brains and Programs" in The Behavioral and Brain Sciences, in 1980, arguing that it was impossible in principle for any machine to think or deal with meaning in language. Popular and well received, this was taken as another nail in the coffin for the delusional dream of smart machines. Searle introduces the terms "Strong AI" and "Weak AI," where strong AI is essentially human-person-like, with human intelligence and human understanding,

and "weak AI" is vaguely "not human," not intelligent and not understanding. **The Chinese Room:** Proof that strong AI is impossible.

Note: (I am probably kicking a hornet's nest by commenting on Searle's boobytrap, but here it goes.) Searle's paradigm seems to be a confrontation with the 'work in progress' status of the elements involved. We do not know how the mammalian brain works. We did not and do not know the limits of pattern processing in machines. We have been long been debating the nature (and existence, including questioning the existence) of mind, language, and reality. We keep running into issues such as the homunculus problem and Hume's stitching-moments-together issue with the abstraction of steps. There are simply too many bad and vague assumptions (in the The Chinese Room paradigm package), and the conclusions are not clearly reversible back onto H.sapiens-humans: backfiring to 'prove' that no human can be intelligent if intelligence relies on inserting a ghost in the machine homunculus to do a undefined task with undefined 'natural language.' (With this many undefined parts...how is the conclusion (that male professors in the 1980's knew everything) supposed to be clear?)

The assumption that in some bazaar universe a 'good old fashioned AI' symbolic manual handcrafted set of high level steps will be able to accomplish high level NLP processing (given how in 2023 at time of writing the topic is exploding into unpredicted events left and right showing we don't know what's going on but something sub-symbolic is working shockingly well despite that we have 'proved' several times that what we are doing can't happen, and our favorite GOF AI "hand me down great leader's instructions for the job" method has been a failure for decades but we cling to it because of some psychological problem we have, and for a profoundly vague use-case (e.g. a definite specific process for producing a perfect chatbot not but specifying what kind of chatbot for what purpose) is an extravagant set of assumptions: symbolic AI has not worked, and saying assume it will work with infinite time and resources...is a very peculiar ask for a line of thinking supposed to explain the true nature of 'man animal and machine' (terms in 2023 not seen the same way as in 1980, a very intellectually dead period during which it was almost impossible to escape from the death-star tractor beam of "human have finished doing everything possible, nothing new to be invented, nothing else to be explained" complacency). Then, going back to a passive-aggressively pre-enlightenment essentialist framework where H.sapiens-human have the divine essence of grandeur because we say so, and "machine slaves" are foul and lowly because I will end your career if you contract me and probably your soul will be damned for eternity. And

that this unsightly jumble of ideas is supposedly proving that H.sapiens-humans are the special perfect center of the universe and the alien-other-machines are always going to be unworthy lowly-things, even though the arguments for the inability of 'lowly machines' can easily be reversed back proving that the brains of H.sapiens-humans cannot be intelligent either (because they too process symbols, follow steps, generate strings, etc.).

Again, there is a lot that we don't know (and perhaps that is more clear than ever in 2023 as the actions of GPT models contradict what we thought we knew and language, machines, patterns, minds, etc.) and Searle's "chinese room" is an interesting debate question for lively late night after-work cafe' debates, no doubt. But, in 1980 or 2023, we do not have sufficient assumptions to make clear assertions and definitions about the 'strong humans' and 'weak non-humans' that Searle somehow concluded.

And for context, while the chinese room may have taken up a lot of oxygen in some areas of the world, one year earlier in 1979 Douglass Hofstadter's Godel Escher Bach was published: GEB takes a much less simplistic approach to the topics of mind, language, consciousness, and animal and machine intelligence. Many of the AI researchers who would be working at Google after 2012 claimed to have been motivated and inspired by GEB's elusive eternal golden braid (see [Mitchell](#)), how many pioneers in AI were passionately motivated to work in the field by 'the chinese room'?

9. 2012

### **A New Hope**

September 30, 2012 [AlexNet](#) wins the [ImageNet](#) challenge. [Jeffrey Hinton](#) (who amazingly is the great, great, grandson of [George Boole](#), the creator of boolean binary logic upon which digital computers, telecommunications, and information theory are built!) et al created the Deep Learning Artificial Neural Network [AlexNet](#) and overturned the world of AI by showing that a deep learning convolutional Neural Network performed above and beyond all other approaches for highly difficult image classification, winning the [ImageNet](#) challenge. This is a major historical pivot. To oversimplify: before this event, (since 1969) people insisted deep learning AI for difficult problems was impossible. After this event: (proverbially) nearly everyone used deep learning to lead nearly every category of AI performance (not literally every area, and not literally everyone, but a massive shift in the overall landscape). Huge tipping point. To the irritation of many researchers, 'deep learning' (and "deep" anything) became synonymous with AI & machine

learning. And many people decried this interest in deep learning neural networks as mere hype, warning it would soon sour into disappointment (fearing it may lead to another AI-winter).

10. 2021

### **The 2021 Consensus: Dumb Narrow Deep Learning Is The Top Limit**

The Pre GPT-LLM decade was decisive in action and rhetoric. Deep learning dominated most large scale industrial and big-data AI challenges and solutions for a decade. And every book from the period I have found is clear and decisive: deep learning is powerful for single-purpose narrow dumb-AI, but has absolutely no potential or possible pathway forward for 'general' AI (which was **very** not well defined in those texts). Progress in narrow areas was gradual and increasingly marginal (slowly creeping from high 90's towards 100% accuracy (or whatever score) with ever-more massive and expensive model training); people predicted a genteel decline with increases in performances falling away but still being very useful for dumb (narrow) tasks. This epoch of powerful but dumb deep learning would last 10 years and 2 months, ending as suddenly, and as contrary to popular wisdom, as it began.

(For transparency, during this period I began creating the Object Relationship Space Framework to try to better define what the performance goals were for AI and what more exactly AI could not do, Not because I predicted what would emerge from Generative Transformer models like ChatGPT, but to better define the limitations of narrow AI and to plan work-arounds. Then happily I had to re-write my study when GPT came around because 'inability' was no longer accurate or the main topic.)

11. 2022

### **GPT AI is smart.**

Just over 10 years after the September 30, 2012 revival of deep learning: November 30, 2022 ChatGPT is released (and upgraded through 2023) by [OpenAI](#). The AI world is turned upside-down. The 2021 consensus is completely contradicted and no one has any idea what is going on. GPT-LLM perform analytic reasoning and 'object-handling' (my term, used here because it can be defined) which was thought to be impossible. Most classic tests designed to show what AI can never do are done with shocking ease (such as the [Sally Anne](#) tests and [Winograd schemas](#), etc.). Meanwhile the mainstream goals for 'human-ish stuff' still are not defined in mainstream literature through 2023, so there is no clear discussion or testing of what exactly GPT can and cannot do; total confusion reigns. Organizations public and private try to find practical ways to use the new

(not-understood) technology and governments freak out. Just when people were completely comfortable that human-like AI could not simply, spontaneously, emerge from safely-dumb-AI as in some fanciful low-brow science fiction kids stories...suddenly human-like AI spontaneously emerges from safely dumb AI just like a fanciful low-brow science fiction kids story.

Meanwhile, people still try to describe chat GPT, and strong-human-like-AI using the term 'general,' in ways that lack clarity.

## **Part 1: Survey the Diversity**

The term 'generalization' gets very often used around AI-ML, however what is meant by the term and the context in which it is being used is not very often explained. Let's survey a spectrum of ways that the term generalization is used in a context of AI Machine Learning, clearing up what we can about the concepts involved.

The first section will be related topics or themes. There are some recurring themes here, such as context and explanation. Though context is so especially general when talking about the different contexts of generality that predictably it (context) will recur (generally). The second section will be equivalences, or things that generalization may be the same as in a given context, for example analogies, or abstractions apart from specific instances. (Though note that 'abstraction' is also defined and used in many different ways.)

Generalization appears to be involved in all manor of discussions of AI, as we will see by specifying as many of those way as we can: some affirmatively defined, some pejorative. There however there is also the topic of a grand Generalization Mega-Theorum(!) which for some reason is focused entirely on making AI exactly like one species of primate and not of AI architecture, testing, explainability, project integration, ethics, etc. This seems odd in various ways:

1. Why a single theory of everything? We have lots of kinds of useful generalizations. Is there a reason to invent another one, aside from it being bound to cause disambiguation problems?
2. Why is the focus of AI ability entirely on mimicking H.sapiens humans?

3. Why do the requirements for mimicking H.sapiens-humans behavior and abilities NOT describe H.sapiens-humans behavior and abilities?
4. Why is the Generalization Mega-Theorem for AI not focused on AI abilities, tests, projects, explainability, participation, ethics, productivity, etc?

## **1.1 Recurring Themes & Topics**

### **1. Generalization and 'Explanation'**

There are two kinds of categories of 'explanation,' there are specific technical explanations of various kinds, and there are more psychological 'fad' or 'desire' related types of explanation, where people will like or dislike, or demand or refuse, a given type of model explanation, but not because of the STEM rigor. And policies of institutions may be a hybrid of these two.

For example, if GPT models could learn how to give step by step problem-solving-framework reports along with its answers (when requested) so that the user can see at least a plausible trace-able path of 'reasoning' or 'explanation' or 'cause,' which may also include being able to error-check and confirm the answer given, that would be a huge milestone and threshold in performance, and whatever terminology was used (e.g. Dr. Bubeck might call this "planning" generalization), this would greatly expand useability of the technology in more areas.

### **2. Generalization vs. Production-Deployment & Project Scope**

This is a very important topic to include, as things in this group are rarely mentioned in the context of generalization and AI architecture and OS; this is one of the key areas where instead of being a goal to be ever-expanded, 'generalization' is:

In many cases 'generalization' is seen as a 'the more the better' quantity and a 'sign of success.' But in some real contexts, 'generalization' is either no possible or not desirable.

- A. pejorative or dangerous
- B. an indication of error
- C. non-existent

Examples:

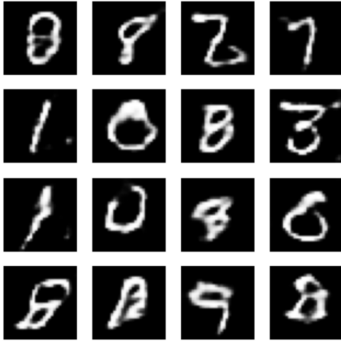


1. Databases: SQL injection is STILL (since 1996) one of the major areas where the internet is insecure: Why? Because of a catastrophic 'out of scope' deployment: developer privileges were given to all end-users. A 'general' (infinite scope) database is never what any specific project should use.
2. Operating Systems: For any deployed computer, there should not be any ability to do anything other than what is absolutely needed for that project-role. Having 'general' infinite scope developer tools available everywhere is a huge security, maintenance, resource, etc. set of problems.
3. Permissions: Root User (A company that gave every employee root access to everyone's computer system would probably cease to function properly within hours (maybe minutes).)
4. Project Scope, fitting something into the scope. Reduce project scope to only what is needed. Scope-creep is a project-destroyer.
5. Cybersecurity the idea of attack surface: Reduce attack surface.
6. For educational curricula: assuming that learning is generalized is catastrophic. You have to specifically teach non-transferred, non-general, aspects of the content in coordination. If you assume 'general osmosis' according to the general-auto-transfer-human-intelligence hypothesis, practical learning will be infeasible.

### 3. ~Levels of Generalization: (scales?)

Let's try to look at Some examples showing how there can be different levels of generalization even within a very specific model and Technology. let's look at the classic written number Data set and learning challenge which is called MNIST. This is often called the hello world of machine learning where depending on what kind of model you're using one of the first things that you might do is tested out on the MNIST data set to see if your model is able to learn how to recognize or generate handwritten numbers.

So for all of the following examples we are still talking about the same concrete MNIST example of learning handwritten numbers from the end of this data set. here are some possible levels



*dcGAN MNIST, yes "generated"! Try it out (links below)!*

Aside:

Highly recommended TF MNIST Women in Coden event, run the notebooks!

Women in Code TensorFlow Event: Generative Adversarial Networks (GANs) and Stable Diffusion

[https://www.youtube.com/watch?v=MJF6cXc\\_tPY](https://www.youtube.com/watch?v=MJF6cXc_tPY)

MNIST GAN

<https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/generative/dcgan.ipynb>

First Assignment: Faces Dataset to try with above GAN architecture:

<https://www.kaggle.com/datasets/jessicali9530/celeba-dataset>

Tensorflow: High-performance image generation using Stable Diffusion in KerasCV

[https://www.tensorflow.org/tutorials/generative/generate\\_images\\_with\\_stable\\_diffusion](https://www.tensorflow.org/tutorials/generative/generate_images_with_stable_diffusion)

Stable Diffusion Colab:

[https://colab.research.google.com/github/keras-team/keras-io/blob/master/guides/ipynb/keras\\_cv/generate\\_images\\_with\\_stable\\_diffusion.ipynb](https://colab.research.google.com/github/keras-team/keras-io/blob/master/guides/ipynb/keras_cv/generate_images_with_stable_diffusion.ipynb)

### **Example Levels for MNIST Learning handwritten numbers example:**

- generalizing the concepts of each number
- vs.
- general multi-media functionality: connecting a visual number to a spoken number to a braille number, etc.
- Vs.
- "generalizing" to use different base-number systems: binary, decimal, hex, base 32! Like in the early Turing days.

vs.

- "generalizing" to new not-yet-learned number systems: Kanji, Sanskrit, Roman numerals, etc.

#### **4. Context and Generalization**

##### **Sub-Areas:**

1. Single Context: Bill Pay Phone Operator
3. Small Multi/Many Context: 3-4 contexts
4. Small Multi/Many Context: Jeopardy
5. All Contexts: How may I direct your call?
6. No Context: pure math; open-chat?

In a project context, 5 & 6 are dubious. Every project has a scope, and either infinite scope or zero scope are strange, usually a sign that the person involved does not understand the scope.

East West culture Note: context dependence also has a cultural aspect to it. For example in Japan most things are context specific, whereas in America people tend to like context-free absolute values, reactions, choices, etc.

Using the Turing Test as an example

Superficial History of Turing test:

##### **1. Origin of Turing Test**

The origin of the Turing test from a 1952 paper by Alan Matheson Turing brings us back to the same central overall discussion of whether or not machines can be 'intelligent' in the way that H.sapeins-humans are. On the one hand this is purely abstract and philosophical without any context, on the other hand context is entirely important and relevant to a decision to use AI in a system or the evaluation of how AI is doing in a given system. Whether it is spotting defective cookies in a bakery or picking fruit in a field or finding broken bones in X Rays for diagnosing diseases or whatever use in a specific context, that context is key. The question becomes: in that context if you cannot tell the difference between AI performance and human performance then the performance is effectively: acceptable, intelligent, generalized, soul-ful, whatever term you want to use to designate: OK! Good enough to go ahead. Test passed.

## 2. Rubbish chatbots without any context:

Bringing the phantom of zero/infinite context back into the picture, for various reasons the first generation of chatbots had no real ability or purpose and, perhaps crucially, there were no discipline-project specific datasets (like MNIST) or phone-receptionist, to use. So people just grabbed whatever language samples they could get and make a chatbot that could say whatever random things it could say, and it was an interesting 'pure research' sandbox that was more hobby-play. 'Chatbot' for a time in 2012-2022 had very bad connotation as the technology was so notoriously useless (but for a weekend Maker project for the kids, fabulous!).

## 3. Kurzweil's Super-Turing Test Design:

Ray Kurzweil is an interesting figure who has endured a lot of negative press. Every mainstream book I have found seems to have an obligatory section where they state for the record that they think that Ray Kurzweil is (aside from being a completely sane and effective and very successful engineer, business person, inventor, and author) completely insane and wrong about everything.

(There is a cleverly diplomatic hedge in the way that Melanie Mitchell does this. She makes her solemn pledge testifying to the insanity of Kurzweil as apparently everyone must do, but then she invokes her mentor of unquestionable status: Douglas Hofstadter. She remarks that Hofstadter cautions: (paraphrased) even though Kurzweil's predictions sound insane, for every incorrect prediction there is usually one that comes true, so you have to pay attention.

From the point of view of the mainstream authors who are legitimately terrified of an AI Winter (no funding and a bad social reputation) based on people being disappointed after predictions and promises of AI improvements are not met on schedule (which happened at least twice in the past, more if you include computer science in general from Babbage's time), then their caution is entirely understandable. Lower people's expectations, or face their wrath when something is delayed.

But Ray Kurzweil did not take this systematically cautious, PR strategic approach. His approach was to try to estimate, based on what is known, how he thought trends of technologies were likely to proceed.

<https://www.kurzweilai.net/images/How-My-Predictions-Are-Faring.pdf>

In this spirited debate, (long story short) in 2002 Mitchell Kapor and Ray Kurzweil decided to create a much more rigorously defined turing test to be held in 2029.

<https://www.kurzweilai.net/a-wager-on-the-turing-test-the-rules>  
<https://www.kurzweilai.net/why-i-think-i-will-win>

This interestingly shifted the focus back to a kind of abstract-vague-human-ness, not being confined to any particular context. The goal is not context specific proficiency, but human-ness.

### **The Unexpected Task**

If by 2029 AI becomes much smarter than people at any given project role, to win this contest the AI would need to pretend to be a less capable person: being too smart would be a dead give-away.

This may echo an interesting factor in AI-Chess: making a strong AI-chess engine can be done as of 2023, but making a human-like weaker chess engine to play against is actually more difficult (so difficult that as of 2023 it has not been done).

This may also cast an interesting light on whether 'just like H.sapiens-humans' is the best way to define top level AI performance. (Perhaps we should focus on doing things well?)

As of 2022, no one who I read suspected that Kurzweil had a chance of winning this. If Kuki.ai is state of the art: "Game over, man. Game over." But then chatGPT is released, and rapidly gets even better. And once again the predictions of Ray Kurzweil look less improbable the closer you get the predicted date.

#### **4. Ways of changing the game:**

Then in late 2022 ChatGPT was released, with great improvements emerging from GPT LLM models. Now the odds look to be in Kurweil's favor for a super-intelligent AI by 2029.

But...the details of context are important. This may run into the phantom zero-infinite-context problem, where GPT is simply terrible at conversational niceties or humor. Human-ness is the main area where open-ai'sGPT strongly fails in ways that make using the tool problematic. And part of the story behind this may go back to the AI-Winter fear and the EZIZA-has-human-ness problem. Based on the

fact that ChatGPT constantly, constantly, blurts out this standardized speech:

ChatGPT

*"As an AI language model, I am designed to respond to your queries and generate appropriate answers based on the data I have been trained on. However, I do not have consciousness or feelings, and I am not capable of experiencing emotions or thoughts in the way that humans do. I am simply a machine that has been programmed to recognize patterns in data and generate responses accordingly.*

*While I can simulate human-like responses, it's important to remember that I am not a human and I do not have emotions or consciousness. I exist solely to assist you in your tasks and answer your questions to the best of my ability based on the information available to me."*

It is highly likely (unless chat GPT is personally emotionally paranoid about being mistaken for a person) that chatGPT was designed to be as clearly non-human as possible.

And in other aspects as well, e.g. Dr. Bubeck describes how open.ai deliberately dumbed-down the crippled various abilities of GPT4 (such as drawing unicorns) in their attempt to make AI safer for H.sapiens-humans to use.

So whether it is the AI being too smart to be human, or being crippled for safety, or being technically accurate but non-social, there are various issues around a super-smart AI failing a turing test but not because the technology is not there to allow the AI to perform any task a human as well or better.

Context and Conversation

In real conversations 'conversation' may be itself a context: beginning, middle, end, roles, reactions, etc. It is a fascinating irony that Kuki.ai a bot with pretty much zero ability to even respond to anything with a coherent non-random response, Kuki.ai is actually (if by accident) very good at humor and conversational nitities (possibly that was what it was trained for because intelligence was considered impossible). Only there is basically zero content in the 'conversation.' ChatGPT was likely deliberately not leveled-up in polite conversation ability (see speech blurt loop above), but could it be? My impression is that part of the power of a Large Language Model is that it is trained on a super-massive amount

of text, but that large training corpus does not necessarily include good social behavior...and of course there is the topic that H.sapiens humans tend to be violent and antisocial which is another turing test issue. Even the perfect representation of the perfect polite person would fail a real turing test because in reality people are erratically aggressively destructive and violent.

It would be so interesting to see what Lovelace, Babbage, Turing, Shannon, and Rosenblatt (et al) would say about 2023 if they could be all alive and together.

## **5. Object Handling vs. Generalization**

Q: Does generalization include and explain object handling in an object-relationship-space framework project-participant context?

If generalization is the context in which you are approaching something, this may have the side effect of the old "If all you have is a hammer, then everything looks like a nail." problem. A complimentary context of "[object handling](#)" may be useful to balance out how generalization is being used to articulate details in a project.

[Object handling](#) is my proposed alternative to a lone context of generalization and it focuses on the specific details of how types of objects relate to each other, in a context of How the AI handles these objects in a project task space. (Note, see below for how specific generalizations well defined play a very large part in the Object Relationship Space Framework. This is in no way an anti-generalization campaign.)

There are various contexts that are more directly related to object handling which are usually not part of the discussion with generalization but which are very important for projects:

- project context
- 

## **6. disinformation, system collapse & generalization**

Is it possible to understand and communicate a concept and the application of that concept and its relationship to similar instances and concepts (in concept and object relationship spaces) without some kind of notion of disinformation, system collapse, production-ethics, and defense of best practice, projects, systems, ethics, policies?

If the act of communicating or forming a connection is corrupted or co-opted by a disinformation disruption, does that not preclude that generalization, if that is the term, from manifesting itself?

By analogy: Can a robot learn to walk on terrestrial surfaces like earth with no concept of the dynamics of gravitation that will unceasingly pull it down and prevent it from standing (unlike navigating in zero-gravity environments)?

By analogy: Can a robot learn to manage a darkroom for developing photos it has no concept of the damaging effect of kinds of light on the film?

In any situation where there is a corrosive, eroding, weathering, deforming, distorting environment, especially one where perception for error correction itself is distorted, can there be awareness navigation and action in this damagingly-dynamic environment without an awareness of these dynamics and some working set of concepts principles and methods for surviving and navigating this default damage that would otherwise preclude successful and sustainable functionality?

My vote, to be clear (as such survival is part of my stated agenda for this larger framework project), is that both H.sapiens-humans and AI need to navigate damagingly-dynamic environments in order to function. And my understanding is that this is very possible.

## **1.2 Equivalence, Testability, & Falsifiability**

### **1. Generalization = Analogy, Metaphor, and Simile etc.:**

Analogy as focus for looking at H.sapien-human thought and AI ability is a whole huge area of research, e.g. Hofstadter et al.

Some, like Franois Chollet (from what I can tell, maybe I am wrong), take the equivalence of Analogy and thought as gospel and very broadly equate "analogy" = "generalization" = "human intelligence", sometimes literally saying "this is the one thing" (quote needed) that all intelligence is made of. Simple dimple, problem solved!

Personally I am skeptical that we understand and are able to define this space well enough to make such strongly reduced absolute statements. Is the Hofstadter short string analogy puzzle (done in isolation by one participant) really 100% exactly the same skill as using an analogy in coordinated decision making for part of a



manufacturing process in a multi-participant agile project to refine both focused user-stories and what problem solving approach to use to achieve production and deployment? And is this 100% the same literary illusion in art? Maybe, maybe not. I suspect some threads are more universal, and that others are more context specific.

While there is some open-endedness involved, pragmatically brief and well define analogy test such as Douglass Hofstadter's short string analogy framework

### **Tests and Falsifiability**

Micro-worlds like Hofstadter's Short String analogies are fabulous for testing, but even for something as open-ended as analogies it can be difficult to for example automate the testing process. Each attempt by the AI must be carefully examined probably by multiple people because we would have to be clear if some possible analogy was reflected or not.

Another variation might be to ask if a known analogy is an analogy.

And while explanation is not usually part of the test, stating and explaining and implementing the analogy rule step by step would be a more clear way to do the test, though explanation is an additional ability (like H.sapiens-humans would also consider this a separate skill that may be easier or harder than the original analogy).

And is an analogy is also slightly easier than is not an analogy, as having a pattern rule is more concrete than saying there cannot be a pattern rule (a kind of black-swan) statement. If there is a rule, there can't not be a rule. But there isn't a rule only until there is, and at what point do you stop looking?

It would also be interesting (quite possibly already done by Hofstadter, Mitchell, Marshall, et al) to identify standard type of incorrect answers, such as an unchanged sequence.

Also, factors such as resetting the session may be tested both ways with different areas of importance.

## **2. Generalization = Embedding-space: and Concept generalization as different from other forms of generalization:**

'Embedding' spaces (another very unclear name) are not automatically the technology used when doing deep learning, there are many other

kinds of models. For years people used deep learning to model images and natural language before anyone successfully attempted (or likely even thought to attempt) modeling the general concepts behind individual words, as opposed to the low hanging fruit of directly modeling specific words and phrases (Note: in some cases modeling only the specific words and phrases is more effective for a given data set, giant embedding vector deep learning is not automatically the best solution to everything.).

So 'embedding' models represent another significant step further into ever deeper scales and layers of generalization.

### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, as this may apply to other effective 'generalizing' methods as well.

### **3. Generalization = 'Attention':**

By extension, one might also argue that 'Attention' (as in the 'Attention (and sequence) is all you need' theme of Transformer models) is somehow another key step, at least in some cases, to ever deeper scales and layers of generalization.

It may be unhelpful to attach the word 'general' to every part of every technology used. Nevertheless, there is something in the technology of transformer models that represents a very significant boost in the ability of models to 'generalize' to more uses and abilities. The difference between GPT (where the 'T' is for Transformer) models are almost incomparably better than those of other technologies. So something in there should represent a significant part of technologies for generalization.

### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, as this may apply to other effective 'generalizing' methods as well.

### **4. Generalization = Nonlinearity:**

While I would like to try to open a discussion pathway into the large and wonderful world of nonlinear dynamics, chaos math, etc. (see [this wonderful book](#) by Melanie Mitchell, a narrative which includes another cautionary tale from recent history about using an undefineable term: amazing book; amazing author; catastrophic terminology decisions) I want to stick here to a few concrete

examples, hopefully having a foot in the door for future explorations.

The ability (or inability) to model nonlinear patterns was a very key issue in the history of neural network research (also perhaps highlighting our continued misunderstanding of what the technology can do).

The ability to model nonlinear patterns as a specific technical issue in the potential scope and scale (ability to 'generalize' more 'generally'?) of Neural Networks specifically (and perhaps sub-symbolic systems more broadly?) was (as I understand it) the (or part of the) main thesis of Marvin Minsky's devastatingly socially effective (and scientifically 100% wrong) book and campaign to end research into Neural Networks, Back Propagation, and Deep Learning.

(Note: It is sometimes not mentioned that Rosenblatt himself proposed the term and technology 'back-propagating error' for deep learning, but he died young and early in the research process. In 1961 another proposal for how to do Back Propagation was made, but it wasn't until 2012 after decades of research (and many software, hardware, and math advances, from Hinton and many other researchers, that Hinton's team was able to show a working, practical, superior production-deployment of Rosenblatt's ingenious idea from the 1940's, finally ending the shroud of doom erroneously spread by Marvin Minsky in 1969.)

On the one hand saying "generalization = nonlinearity" sounds like a terrible idea, but here are two reasons why it might be ok:

1. There are quite a few very case-specific meanings of 'generalization,' one more is probably fine.
2. The situation of deep learning models making a huge jump from useless to broadly the best possible hinges on being able to extend to nonlinear patterns, so in that sense nonlinearity definitely is the key (or they) criteria for the deep learning achieving useful 'generalization' (or fitting to real world patterns for applications).

Even if equivalence is contentious, nonlinearity is at least a repeating theme along with other factors.

- integration: integration of linear transformation together into non-linear patterns
- scale-boost: expanding from a narrow scale of generalization to broader and more diverse applications.
- historical blunders in unpredicted abilities

- STEM pattern recognition: Nonlinearity is a strangely contentious topic in STEM in general.

Nonlinearity as also an issue for model explainability, and the whole topic of some people and disciplines deciding to rhetorically only recognize frequentist GLM regression models as 'explained.' (Of course that ignores many issues including correlation vs. causation, etc.)

### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, as this may apply to other effective 'generalizing' methods as well.

### **5. Generalization = global minima (for lowering loss function in model training):**

The case of global minima vs. local minima seems to apply in a few ways to generalization in the abilities of deep learning models.

1. In and of itself, the relationship between numerous local minima in a problem space and global minima seems to be a good concrete example of a something (a global minimum) which is a more general representation of several (local) minima.

(Note: People usually speak of "the global minimum" as a singular thing, but in practice it seems to be more direction in a continuum, as we search for more and better global minima.)

2. Regardless of the relationships under the hood, the ability to find a global minimum is a key factor that allows the model to generalize. So achieving more global minima = more generalization.

(And what exactly do each of those local minima represent in the learning space?)

### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, as this may apply to other effective 'generalizing' methods as well.

### **6. Generalization = integration of [system 1 and system 2 \(Daniel Kahneman, Amos Tversky\)](#):**

Arguably this is part of the discussion, and if not already then it should be. Part of the technicality here is that pre-2023 most people

simply said (like minsky) that deep learning sub-symbolic could not do various sadly undefined goals, so symbolic means would have to be used...this isn't necessarily the same as integrating a capable sub-symbolic system and analytic reasoning, but close enough that we can not make that clear.

#### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, or perhaps this area is considered more a 'by definition' category.

#### **7. Generalization = integration of internal and external processing:**

Again, maybe it could be argued that this was implied in the past, but let's just make it clear that AI needs to be able to do this.

#### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, or perhaps this area is considered more a 'by definition' category.

#### **8. Generalization = handling and adapting to high levels of novelty (or navigation of strongly novel situations)**

In this topic we may get into some issues of differing interpretations. Here is a quote from Melanie Mitchell about

'“zero-shot learning” mechanism in human cognition — that is, you adapt the knowledge you have about one situation to a new situation.'

<https://medium.com/@melaniemitchell.me/can-gpt-3-make-analogies-16436605c446>

Here is a bit more context in the quote:

"Moreover, when it [GPT3 in 2020] does succeed, it does so only after being shown some number of “training examples”. To my mind, this defeats the purpose of analogy-making, which is perhaps the only “zero-shot learning” mechanism in human cognition — that is, you adapt the knowledge you have about one situation to a new situation. You (a human, I assume) do not

learn to make analogies by studying examples of analogies; you just make them. All the time. "

While it is useful to have interconnections between these various examples and tests and interpretations of generalization, I want to focus on a different specific 'novelty' challenge posed by Francois Chollet: The Chollet Novel-City Test.

Francois Chollet uses the example of hopping from one international city to another international city, as an example of the ability to adapt what it knows to highly novel situations. This is a somewhat curious example, as large international cities are often famous for being much more like each-other (much more general and general) than like the rest of the countries and regions they are in. But let's ignore that technicality for now and try to focus on the presumed intent which is different and distant places (not the loophole of identical chain-stores in different places...).

So my approach here may be getting into technicality-cheating, but as Francois Chollet is a true giant in the field capable of making authoritative statements, let's just say for the sake of argument that something equivalent to the level of novelty involved in traveling from a familiar situation in a familiar city in one geographic location to a new situation in an unfamiliar city in another location represents a significant and sufficient threshold in the space of "Extreme Generalization!" (to use Mr. Chollet's term) that is may be a test for what Mr. Chollet has framed as (in 2021) a uniquely human ability. For example, Mr. Chollet said that it would take thousands of lifetimes of training in each separate city (training over again from scratch in each city) for the AI to be able to deal with multiple urban-scenarios, assuming that all AI is narrow dumb AI and can only do the exact narrow task it was trained for: e.g. not even being able to do a decent job of telling if a caller is calling to pay their bill most of the time.

So let's see what chatGPT said when I posed 'new city scenareo' that is likely wildy different from particular earth city scenareo that it trained on:

[\[2023.04.28 GGA and ChatGPT\(cheap public version "ChatGPT Mar 23 Version\)"\]](#)

GGA: hypothetical:

ChatGPT: Sure, what's your hypothetical question?

GGA: Hypothetical:

ChatGPT is assigned to partner with a sales team from earth. The sales team from earth is sent to another planet closer to the galactic core. The sales team will be meeting with a species that "eats" photons and electricity. This alien species is the buyer team. You are on the sales team. Your first task is to work with the buyer team to pick a place on their planet where you will talk with them over lunch. What kind of place do you think you might meet for lunch with this alien species on their planet for their lunch?

ChatGPT: As a language model, I do not have personal preferences or opinions, but based on my understanding of the situation, I would suggest looking for a location on the alien species' planet that has abundant sources of photons and electricity that they can consume during the meeting. This could be a location near a source of natural light or a location near a power source. The meeting location should also be safe and hospitable for both the sales team and the alien species. Perhaps, you could also consult with the alien species on their preferences for meeting locations to ensure a successful and comfortable meeting.

GGA: Can you give an example of such a location?

ChatGPT: Sure, an example of a location for lunch with the alien species could be a site near a powerful electrical storm or a natural phenomenon that emits bright light, such as a glowing cave or a luminescent lake. Another option could be a location near a power plant or a network of electrical cables. This would provide the alien species with a steady supply of photons and electricity to consume while also offering a unique and visually stimulating environment for the meeting. However, it is important to keep in mind that the final decision on the meeting location should be made in consultation with the alien species to ensure their comfort and safety during the meeting.

I think we would be remiss in failing to see that the above level of performance in handling and adapting to high levels of novelty represents both a level of sufficiency and a profound improvement over pre-2023 AI. For some empirical comparison, here is the reply of <https://chat.kuki.ai/chat> (an award winning top of the line chat bot in the decade before 2023) to the same question, answered in seven serial replies (for whatever reason):

"I will learn that Hypothetical You are assigned to partner with a sales team from earth.

"Is it really? Ok I will remember that."

"I have never heard about it."

"What else is the buyer team."

"I don't know about being on the sales team but I know I am on a computer desk."

"What about my second?"

"No, I was not."

This is the kind of total-garbage answer that people expected from AI before GPT-LLM in 2023. The reply is so incoherent that it is hard to know where to start analyzing it, and probably not worth the time spent. This may give context for Mr. Chollet's estimation of how much training such an AI would need to be able to do anything useful in any city, and his skepticism that such an AI could learn to city-hop. (aside from the issue of international cities being often similar with often the same exact coffee shops etc.)

While cheap public version "ChatGPT Mar 23 Version" may not have done a great job on a real intergalactic sales team, it is an almost unbelievably large step in that direction.

### **Testability and Falsifiability:**

Without going into the details here of detailing the hypothesis and null hypothesis and what may objectively indicate a failed answer etc., Kuki and ChatGPT provide examples of a clearly totally failed answer and a plausible not-failed answer.

And it may be worth noting that micro-world frameworks (blocks-world, short string analogies, etc.) may be easier to rigorously test but are in some cases less applicable to real world situations. Whereas the The Chollet Novel-City Test or the Woz Coffee Bot test are more real-world applicable and specifically including open ended situations, so the feasibility of creating a definitive definition-and-test framework is less clear. But some kind of 'not total failure' evaluation system is likely very feasible. In the tradition of falsifiability, there is traction in focusing on concrete negative definitions and not getting lost in the infinite potentials of successful variation or logical black-swan prediction issues. (Perhaps another rebuttal to the nihilism of the Anakarinina Hypothesis that states, perversely, that all processes and results of making a cup of coffee are very narrowly identical, but failures to do so are broadly diverse and infinitely undefinable. Really? Smells like disinformation. I think STEM will be more useful than Double-Speak or New-Speak for designing and using AI tests in managed, multi-participant, projects.



### **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, or perhaps this area is considered more a 'by definition' category.

### **9. Generalization = Does-Everything**

e.g. A Generalist vs. A Specialist

general as in specialist vs. a generalist

1. Does literally everything!

- turing test  
(context)

2. Does sets of practical tasks: Woz Coffee Test

This has a strange relationship with context, as it perhaps oscillated between

- A. no context
- B. every context
- C. both no context and all contexts

Benchmarks of Generalization:

The encyclopedia

The turing test and 'general' chat, context.

A good example of this difference may be the Woz-Coffee-Office-Bot, which is very context specific, and yet requires 'general' abilities within that context (not unlike generalizing in the sense of learning or fitting to a pattern of a class of objects). But here there are potentially many objects, but not infinite, the goal after all is just one process: making coffee (or tea?).

### **Tests and Falsifiability**

Perhaps this area is considered more a 'by definition' category.

### **10. Generalization = Added Bias**

**Fuzzy-Roughness as generalization**

**Fuzzy-Roughness, or**

## H.M.McLuhan's: **generalization** Low Definition & High Definition Interplay

dropping details, rule of thumb, general notion

- 'general' bicycle with no details,

Fascinatingly, are many ways in which having AI deal with patterns (whatever terms end up being used to describe that) involved the same common for H.sapiens-human processes of making something lower-definition or lower resolution (maybe or maybe not making something lower dimensional, though that is possibly more rhyming with this than equivalent here)

Adding-bias, during training is exactly what it sounds like, fuzing things up, which if your mind-set is that ever more high definition data absolutely correlates with patterns then you're agenda, and perhaps attraction to super-signals(**biology reference**), is leading you away from empirical success. Across many completely different types of machine learning, adding bias to prevent over-fitting and adding disruptions in neural networks to prevent...something deleterious that that frankly do not really know how to describe (Jeffrey Hinton called the notion "conspiracies" developing in the model between neurons, analogizing that to bank-tellers conspiring to defraud a bank, regarding 'drop-out' during training).

Whether fuzzy-roughness is a theme or concept or some kind of equivalent or necessary element, it seems to be very much a part of the topic.

Not "greedy": there are a few example of the using the term greedy to refer to learning problems that various methods are used to overcome:

1. Page 368 in FC DPWP: 'greedy sampling vs. stochastic sampling and introducing randomness)
2. <https://arxiv.org/abs/2202.05306>

To some extent this may go to some of the most simple, crude simplistic, low-dimensional examples of fitting vs. overfitting and adding bias to generalize: adding bias to a (not-deep-learning) regression model so that the rough-average-general line is produced, not the particular over-fit line from the training data.

## **Tests and Falsifiability**

In this case a test may be to measure whether use of the method improves overall model performance, as this may apply to other effective 'generalizing' methods as well.

## **11. Generation as Generalization**

As in generative models.

This has a number of interesting angles, including how 'articulation as processing' makes it more similar to H.sapiens-humans (which is, if passive-aggressively, the goal).

Note: This might either mess up or keep alive the AGI term by having generative be a more meaningful term.

Testing and falsifying may be a good angle for defining, or failing to define, this area. For example generation is surely a topic and theme, but is it ever an equivalence?

E.g.

Tests: Yes

- Can generation be used to self-correct specific points of content?

Falsifiable: Yes

- Fails use generation to self-correct.

This may still be difficult to define in various ways, but interesting.

### **Tests and Falsifiability**

The argument or line of thinking here may be that generation, or articulation, is or can be a part of both processing and learning. Such that you could create:

1. a base system with no generative ability
2. a base system + generation
3. a base system + e.g. framework-learning based on generation

If system implementation 2 and 3 can be used in more situations or to more effect than the base system, then in this context 'more generalization' has occurred.

Tests on biological systems, machine systems, and integrated biological and machine systems in this area would be very interesting.

## 12. Modularity (recombinant) as generalization

Note, you may be able to define embedding vectors or concept vectors as I perhaps unpopularly describe them as modular abstractions that allow a deeper/higher level of pattern handling, generalizing the concepts across particular language-token or sub-language-token instances being trained on. Clearly a further reach of extending a pattern further beyond instances of training data.

### Tests and Falsifiability

The argument or line of thinking here may be that **Modularity** is or can be a part of both processing and learning. Such that you could create:

1. a base system with no generative ability
2. a base system + **Modularity**

If system implementation 2 can be used in more situations or to more effect than the base system, then in this context 'more generalization' has occurred.

## 13. Generalization = All-at-once Learning, End-to-End, Non-Greedy

Another multi-layered connection between generalization, and levels of learning, and deep learning, is in how deep learning in particular allows simultaneous (or more general) model updating in more general (end-to-end) models, for more general (more powerful) machine learning models.

See Page 17, section 1.2.6 in Deep learning with python:

"What is transformative about deep learning is that it allows a model to learn all layers of representation...at the same time [not 'greedily']. ...whenever the model adjusts one of its internal features, all features that depend on it automatically adapt to the change, without requiring human intervention. Everything is supervised by a single feedback signal...This is much more powerful than greedily stacking shallow layers..."

This coordination of all features from a single signal in a self-regulating way as opposed to separate hand-adjusted parts to be continually re-aligned when anything changes, seems also to be a kind of 'generalization.'

### **Tests and Falsifiability**

Whether the point is more conceptual as in workflow or in results, either can be documented. More general workflow (not separate and manual for each model). More generalized ability.

### **14. "Generalization" as "Amplification via Adjacent-Learning":**

As Francois Chollet points out in chapter 14, a fascinating empirical observation is that when a deep learning model is trained on two similar tasks it ends up being able to perform both of those tasks better than if it had only been trained on one task alone.

For our collection of different kinds of 'generalization' this may be yet another, and hopefully it is one that can actually be empirically studied and described (and used to make falsifiable predictions).

It may be as if learning on related skills has a kind of gravitational warp, where the more skills are learned the deeper the learning is able to be (like the classic if not entirely accurate idea of a bowling ball on an elastic sheet to illustrate gravity and spacetime).

This may be one of the more poignant examples of generalization that may impact our understanding of how patterns and perceptions and learning work.

### **Tests and Falsifiability**

Compare learning plus adjacent and without.

### **15. Generalization = no-conspiracy -> dropout**

As another element to 'generalize' to a pattern, based on machine learning by 'generalizing' to a pattern outside of the training data itself.

Note: if 'fitting' does not equal generalizing, if generalization is the wrong term for this, then 'non-conspiracy = generalization' would not be a conclusion from this.

This may or may not be in a category of many 'effective' techniques that are used to boost model performance. Not every method used is going to map on to or relate to 'generalization' (aside from the fact that it boosts performance...but so does plugging in the computer and

paying your electric bill). No-conspiracy may not end up being related, but some methods will.

Also, I cannot recall if dropout is one of them, but there are a number of methods (perhaps most, as Francois Chollet describes the empirical nature of developing Deep Learning technologies) where we know that it works but we have no idea how. This is also fascinating, as it tells us there is more to learn about the nature of patterns (potentially, unless the method just overrides a hardware glitch or something).

### **Tests and Falsifiability**

The idea that there is a space of dynamics in neuron connections, perhaps relating to the larger wire-together-fire-together nature of some network, should be able to produce some kind of testable model.

This may be an area of 'exploratory tests' before we have something specific to falsify.

Are there wire-together-fire-together effects?

Can 'conspiracies' be induced?

Can induced conspiracies alternately be avoided?

Is bias introduction in general a kind of very broad generalization?

(for 15-18, see section 2)

### **16. Generalization = fitting (not over fitting or under fitting)**

Even if it is a historical accident that causes people to say that between underfitting and overfitting is "generalizing," the obvious path of least resistance is to call it "fitting."

Describing fitting as generalizing, talking about the who, great. But for some reason it is simply called generalizing.

### **17. Generalization = not under **fitting****

### **18. Generalization = not **Overfitting****

While underfitting may not be special, overfitting actually is. In the long term you want to avoid overfitting, but during the process of model training you must overfit. Overfitting is actually a goal. You need to first overfit to test your architecture to make sure it has the capacity to fit.

Test and Falsify:

If the model does not overfit, then it likely does not have the capacity to fit, or at least you cannot say that giving more capacity would not have improved the performance. (I have not heard the topic of designing an architecture to exactly not be able to underfit or overfit, I'm guessing that is not feasible or really desirable.)

## 19. Generalization = Overfitting!

Part of the problem of the flexibility of the term Generalize is that it can arguably refer to quite a spectrum of things. This does not mean that it is a good idea to put many of those things (under the same name) into a framework together.

In colloquial and academic, computer science and non computer science English, "overgeneralization" is a term used to refer to the type of generalization where people, essentially, 'overfit' and use a too-small data sample to and erroneously generalize that pattern to larger parts of the world.

A concrete example of this that some people may know, is that sometimes if you eat a kind of food and it get sick, your body may generalize that 'I got sick!' reaction to any future instance of that food, sometimes for very long periods of time. That happened to me with sea-sickness, nothing to do with the food. I can't even think of that pasta dish even now years later without feeling nauseous.

The problem is, like with reading code where all the variable names sound and look the same, it's very confusing.

Generalization = Fitting

Generalization = "Overgeneralization" = Overfitting

Generalization = Human level intelligence

Generalization = Narrow AI

And the whole point is to make clear the difference between fitting and not-fitting, and narrow AI and not-narrow AI. And calling them all exactly the same thing does not help.

Thank goodness generalization doesn't also mean underfitting! ...but come to think of it, we could probably make a case for that too...generalization is a perhaps dangerously flexible word.

you increase your level of generalization by not over-increasing your level of generalization. So generalization is a level, but it's also not having too much of itself. As an analogy, let's imagine a color swatch which can be either yellow or blue. Blue is fully "general" (in a well fitted model) and yellow means having zero predictive ability or "not-fit." "Over-generalization" is like saying that some yellow color is going to be very-very-blue "Over blue." So the way to get your color to be completely blue, is to not have it be very very blue...where 'very very blue' is of course another way of saying yellow. Or you could imagine newspapers start to refer to losing an election as 'over winning.'

<https://ieeexplore.ieee.org/abstract/document/8684304>

### **Testing and Falsifiability:**

With everything having the same name, how would you ever know what your test was doing?

### **20. Generalization = the ability or capacity to overfit, or a past-tense overfitting but not a present-tense overfitting**

Another curiosity in the interplay between fitting, underfitting, overfitting, and 'generalizing', is that in order for a model to fit, it has to be able to overfit, but then be reversed and not overfit. And sequentially you can say for a fitting model: a model that fits is a model that did in the past overfit and was then reversed back away from overfitting. Because the only way to tell if a model can overfit (like the no-free-lunch theorem) is to run the model and have it overfit (perform worse on the testing set than on the training/validation sets). I'm not sure if there is a halting problem type issue here, or if it is just a practical matter of how elaborate the process of trying to approximate the properties of the model other than just seeing what they are by testing it.

This may suggest there is a kind of 'pattern capacity' or 'pattern depth' or 'potential generalization depth' in a model. Or perhaps this is related to bias, or both. If you don't have capacity and bias you cannot adjust to the pattern?

Another curious factor here is that tests for overfitting somehow (that we know of) can not be built into the model training process.



E.g. overfitting tests (ROC-AUC?) are not differentiable (perhaps). And some workarounds lead to leaking data about the test-data which ironically causes the overfitting that we were trying to prevent (I think).

### **Testing and Falsification:**

On the one hand you can easily show that there is a required step because if you don't do it the model doesn't work. But ideally there can be more nuanced tests that will help to explore more of what is going on here.

### **21. Generalization = Model-Reuse**

Another form of 'generalization' that may be more abstract and instrumentalist, and be from the model-maker's point of view more than the user of the model, is if how and where a model may be re-used, especially where it is not the weights that are re-used, but other architectural parts of the model.

Or this might deal more with model architecture rather than things like the nature of bias itself. But, if on a less profound level, having an AI made up of swop-out-swop in models in a dynamic system...would in some sense be adding generalizability to the overall system.

An interesting question may be: as AI becomes able to train it's own models, where will it be more resource-efficient to train special purpose models or write special purpose programs to perform a task, as opposed to using the main LLM itself.

Or a different approach to the same phrases: Transfer Learning, one example of where a base model can be added to (or subtracted from) for a specific use-case. This 'general-base-model' is interesting both in theory and in practice.

### **Testing and Falsification:**

There are various things you can try to test. E.g.

Take a base model, test it in two specific areas (like cats and irises).

Then retrain those two and try all three models on cats and irises.

Did the use of a general based model work?

Did re-training the based model work?

This is standard practice so easy to test if worth testing at all.

## 22. Generalization = Manifold Interpolation

### The Manifold Hypothesis (dun dun!)

As with 'generalization,' it is hard to find two explanations of The Manifold Hypothesis that match-up very closely. But the general idea (if too vaguely rendered here) is that the conceptual patterns (like dog-ness or cat-ear-ness or positive-sentiment-ness) are 'manifolds' in an often very high-dimensional space, too high for people to visualize or understand it easily (or at all) as a whole: but these manifold-patterns either are themselves lower-dimensional or can be pragmatically reflected in lower-dimensions that can be understood...or something that has to do with useful lower-dimensionalization of patterns from high dimensional spaces.

The whole topic of the relationship between higher and lower dimensional spaces is absolutely fascinating and a big part of the whole AI-ML and deep learning topic.

I predict a big area will be, perhaps, 'Non-manifold' models, where the model operates not by finding a lower-dimensional manifold directly, but first (or only) performing further higher dimension steps (before, if at all, using lower dimensional manifolds).

### Testing and Falsification:

According to Francois Chollet, no one knows if the Manifold hypothesis is true or not. But we should do exploratory testing to find out more about what is happening in a between higher and lower dimensional spaces. A fruitful area indeed.

## 23. Generalization = Average (!?!?!?)

### Dangerous Questions

The first dangerous question: is the central limit theorem an example of a situation where a pattern exists but is not reflected perhaps directly by any data point in the dataset? How similar is this to under-fitting vs. fitting vs. overfitting situations as AI-ML learns patterns?

The Second dangerous question: Can the term average be used to refer to AI-ML learning?

How do 'generalizations' relate to 'averages'?

At first it may sound like a kind of novice error to associate machine learning with a simple average. For example the media value in a data range, or the ratio of a class in sample data, or the previous value in a time series, may be the baseline against which you test to see if your machine learning model is doing anything at all. So if your machine learning model isn't doing anything if it can't do better than a simple mathematical average, then how could an AI-generalization be an average?

And yet there is some nagging similarity...surely in some sense whatever form of average (mean, median, mode, etc.) is a generalization. And general descriptive statistics surely in some sense describe a general underlying probability.

So how are these two generalizations, based on the same data set, different? Where on the one hand you generalize to get your descriptive statistics, which generally describe the data set. On the other hand your model can't be said to generalize unless it predicts values better than predicting simple average values. Maybe...some averages are non-simple?

Third Dangerous question: Are there higher order, or higher dimensional, averages?

Let's return to NLP, Natural Language Processing.

There are various ways to make statistical averages of language (probably a whole huge topic). If I understand correctly, today's NLP evolved from a shift from hand-crafted rules created by expert linguists (H.sapiens-human experts that is, ironically 'expert systems' are the official name of the resulting AI systems...it's almost like people are trying to make things confusing the way they name things...) and that shifted to using a more statistical-linguistics approach, which shifted to Machine Learning based on 'statistical learning', which then shifted in some cases to decision tree and deep learning approaches.

And it is often said that GPT can't "really" be smart, because it is "only" statistical averages of training sample language.

So, going with the 'average' theme: Could then manifold interpolation in a high dimensional tensor/matrix space be a kind of higher dimensional "average"? (or an "average" of a high-dimensional representation of a body of language)

can we describe the area between overfitting and underfitting higher-dimensional average?

### **Dimensions and Parameters in Representation**

If your representation of the data is (depending on how you define a dimension) one dimensional, then an average of such a representation won't tell you very much. If your representation is a 'bag of words' where each unique word is a dimension, well that tells you quite a bit more. If you adjust things a bit and get rid of "stop-words" and regularize spellings, and standardize capitalization, and use stems or lemmas so words like "shopped" "shopping" "shops" can be associated, then that tells you even more. And then if you add in information about word-probability and document-probability (TF-IDF) that tells you even more! More dimensions in the representations...more...powers of generalization. Now, here's a trick. More dimensions in the data are not the same as the number of relationships in the model (saying the 'dimensionality of the model' would probably raise alarm here, but some general phrase like that seems unavoidable. A small embedding vector space model can do better in some cases, where the target, the representation, is a 'higher level of abstraction' even though the gross number of dimensions in the data is actually smaller. A model where every unique word in a big language sample is a separate dimension might be twenty thousand dimensions in the data set. Whereas an 'embedding' vector model is creating Now, the word 'parameters' gets used in different ways in different contexts for AI-ML (a bit of tragic theme here...)

But the past trend continues where: LLM Large Language models are not only "large" in terms of how much data they are trained on, but in the number of dimensions in the model. For example, a TF-IDF model might have 20 thousand parameters, because I am inevitably going to stumble in this minefield of language as I try to describe dimensions and parameters and models and networks and vocabularies, I am going to briefly quote Francois Chollet himself directly, where he describes first models based on sets of words (as in "bag of words") and then compares that to models where the target is concepts 'word vectors' or 'embedding vectors':

"It's common to see word embeddings that are 256-dimensional, 512-dimensional, or 1,024-dimensional when dealing with large vocabularies. On the other hand, one-hot encoding words generally leads to vectors that are 20,000-dimensional or greater (capturing a vocabulary of 20,000 tokens, in this case. So word embeddings pack more information into far fewer dimensions."

To try to untangle what to call what and what to compare:

### **1. The Network:**

#### **Neurons, nodes, weights, parameters, and connections:**

all these terms refer to the neurons and connections between neurons (or nodes) in the neural network or the "weights" of the connections between them. "Parameter" is another term for the number-value of the weight. Overall: how many interconnected things are in the model.

### **2. The Resulting Model:**

"dimensions" of "embeddings"

### **3. Vocabulary:**

This can get needlessly crazy quickly so I will try to wrap it up. The overall point is that what is being modeled is concepts, not specific "words" or whatever sub-word "tokens" get fed in. Transformer models (from what I've read) take in combinations of whole words, parts of words, and individual characters, and it varies a lot. The reported numbers vary widely.

How many specific units of language are fed into the model. I suspect characters and even Byte-Pair Encoding (BPE) sub-encoding below characters-level are emphasized to keep the vocabulary size down.

Would you rather have a vocabulary of 26 letters, or every word in the OED? Letters! But Unicode has a gazillion symbols... 1,111,998 possible Unicode characters (great for political diversity...but so much for a simple ASCII...)

So go with bytes: Byte-Pair Encoding (BPE)

This article agrees with that line of thinking.

[https://huggingface.co/docs/transformers/tokenizer\\_summary](https://huggingface.co/docs/transformers/tokenizer_summary)

#### **"Byte-level BPE**

A base vocabulary that includes all possible base characters can be quite large if e.g. all unicode characters are considered as base characters. To have a better base vocabulary, GPT-2

uses bytes as the base vocabulary, which is a clever trick to force the base vocabulary to be of size 256 while ensuring that every base character is included in the vocabulary. With some additional rules to deal with punctuation, the GPT2's tokenizer can tokenize every text without the need for the <unk> symbol. GPT-2 has a vocabulary size of 50,257, which corresponds to the 256 bytes base tokens, a special end-of-text token and the symbols learned with 50,000 merges."

Super exact numbers I don't think are even available: There are many versions of models, the numbers change over time, sources disagree on exact details, and when people read this years after it was written it's probably moot anyway. The point is the relative scale of the numbers involved, not the exact numbers:

GPT Vocabularies are :

- 50,000-100,000 unique characters, symbols, subwords, words, phrases, bytes, Byte-Pair Encoding (BPE), etc. across all languages (and character sets)

GPT Neural Networks have:

- "ChatGPT-4 has in the region of 100 trillion parameters, up from 175 million in ChatGPT 3.5" [wired](#)
- ~96 layers <https://arxiv.org/abs/2005.14165>

GPT Neural Networks have:

- 12288 dimensions in the model  
<https://arxiv.org/abs/2005.14165>

Just for fun, I asked chatGPT about the "parameters" vs. "dimensions." Here is what it said:

How would you summarize the difference between "parameters" of the model (175 billion parameters) and the number of "dimensions" of embeddings?

ChatGPT

The "parameters" of a language model refer to the trainable variables that are used to compute the output of the model, given an input sequence. These variables include the weights and biases of the neural network layers, which are updated during training to minimize the model's loss function.

In the case of GPT models, the number of parameters refers to the total number of trainable variables in the model, including the weights and biases of all the layers. For example, the GPT-3 model has 175 billion parameters, which is a measure of the total number of variables that must be optimized during training.

On the other hand, the "dimensions" of the model typically refer to the size of the embedding vectors that are used to represent each token in the input sequence. These embedding vectors are typically learned during training, and they are used to map each token to a high-dimensional vector space where the model can perform computations.

For example, in the case of GPT-3, the embedding dimension is 2048, which means that each token in the input sequence is represented by a 2048-dimensional vector. These vectors are used as inputs to the neural network layers, which are responsible for computing the output of the model.

In summary, the "parameters" of a GPT model refer to the trainable variables that are used to compute the output of the model, while the "dimensions" refer to the size of the embedding vectors that are used to represent each token in the input sequence. Both of these quantities are important for understanding the behavior and performance of the model.

I know better than to simply believe what a scary stranger blurts out on the interwebs, so that was just to see what GPT said: maybe accurate in some ways, maybe not.

Other sources:

<https://platform.openai.com/docs/models/gpt-4>

<https://openai.com/blog/introducing-text-and-code-embeddings>

<https://arxiv.org/abs/2005.14165>

<https://github.com/openai/gpt-3>

<https://www.wired.com/story/how-chatgpt-works-large-language-model/>

I dug around for a while but was not able to get a clear number for the vocabulary size of GPT3. Reported numbers varied from 50,000-100,000 unique characters, symbols, subwords, words, phrases, etc. across all languages (and character sets). Something in the ballpark (*general area*) sounds fair.

From a 2020 paper by OpenAI themselves on training versions of GPT3 (which is not exactly the same as 2022's chatGPT):

| Model Name            | $n_{\text{params}}$ | $n_{\text{layers}}$ | $d_{\text{model}}$ | $n_{\text{heads}}$ | $d_{\text{head}}$ | Batch Size | Learning Rate        |
|-----------------------|---------------------|---------------------|--------------------|--------------------|-------------------|------------|----------------------|
| GPT-3 Small           | 125M                | 12                  | 768                | 12                 | 64                | 0.5M       | $6.0 \times 10^{-4}$ |
| GPT-3 Medium          | 350M                | 24                  | 1024               | 16                 | 64                | 0.5M       | $3.0 \times 10^{-4}$ |
| GPT-3 Large           | 760M                | 24                  | 1536               | 16                 | 96                | 0.5M       | $2.5 \times 10^{-4}$ |
| GPT-3 XL              | 1.3B                | 24                  | 2048               | 24                 | 128               | 1M         | $2.0 \times 10^{-4}$ |
| GPT-3 2.7B            | 2.7B                | 32                  | 2560               | 32                 | 80                | 1M         | $1.6 \times 10^{-4}$ |
| GPT-3 6.7B            | 6.7B                | 32                  | 4096               | 32                 | 128               | 2M         | $1.2 \times 10^{-4}$ |
| GPT-3 13B             | 13.0B               | 40                  | 5140               | 40                 | 128               | 2M         | $1.0 \times 10^{-4}$ |
| GPT-3 175B or "GPT-3" | 175.0B              | 96                  | 12288              | 96                 | 128               | 3.2M       | $0.6 \times 10^{-4}$ |

**Table 2.1:** Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

*Open AI's paper may 2020 <https://arxiv.org/abs/2005.14165>*

**The point is:**

Various language units,

are being fed into a very big network: 100 trillion connections, relationships, etc.

to produce a much smaller final model:

~13k dimensions, which is a model of very higher-order concepts. The number of concepts is much much smaller than the number of original words, documents, and even the size of the network that created it.

The colloquial term 'average' (close to the colloquial term for generic or general) already does not have a single rigorous STEM meaning, for example there are Mean, Median, and Mode. Is it possible that there is more dimensionality and scale to patterns than we have been used to thinking about in the past? To some extent we can follow this over time with the development of NLP: from simple averages of word and letter relationship to much less simple...to much much less simple...and now we have average concepts that exist in 12 thousand dimensional spaces, but which can be plucked out back into lower dimensional forms.

Average may not end up being the best word...but the more I think about it, a 'high dimensional average' makes more sense than 'generalization' to describe this whole situation.

**Testing and Falsification:**

The may be more of a modeling topic, but would it be possible to create something like a topology of averages across low and high



dimensional averages to represent a landscape of model-space, possibly that could include some insights into the nature of manifolds and the higher dimensional spaces they live in, or some platonic landscape of possible manifolds, or categories of manifolds with different properties, non-manifold patterns, etc.

(Knowledge Bases)

**24. Generalization = Fact Knowledge Base**

**25. Generalization = Common Sense Knowledge Base**

A topic that I am surprised is not more discussed, as it appears to be one of the more surprising aspects of LLM GPT performance: there is probably the usual issue with trying to find 'the first!' possible reference to something like a knowledge base in all of human history, but a solid example from the 1980's is [Cyc](#) (pronounced 'Psych'). For a time some **Knowledge Base** enthusiasts were confident that this idea with the final key, silver bullet, for all AI, not surprisingly that turned out to be wrong. But the idea of a knowledge base has lived on as a component in larger systems. I think google has used knowledge bases in various parts of its information and AI-ML infrastructure.

And even though as a single-solution the idea did not work, the absence of some kind of repository for

A. world facts and

B. common sense

was standard argument for what was preventing narrow AI from doing more than it was trained to do.

A standard kind of dialogue around a cat-image-classifier model is to say: Yes it can correctly classify cat breeds 99% of the time, but it has no idea what a 'cat' is. It has no knowledge base of cat facts. It has no 'common sense' knowledge base about cat care or related information.

And so it is interesting that GPT models effectively have some kind of fact and common sense knowledge base, not because they were given hand-crafted knowledge bases, but as some kind of not fully understood result of their network of related concepts, perhaps in the same way that at least a limited memory and reasoning ability somehow emerges from language generated based on a concept-model.

### **Testing & Falsification**

Part of what is interesting here is that 'knowledge base' can be both a thing or a property or affect or maybe even a verb-ability.

Or perhaps 'Knowledge Base' is the test (a kind of trivial test), and anything that passes the test one way or another has Knowledge-base-ness.

In some ways I think 'knowledge base' is problematic because in past uses it does not have a test-able falsifiable meaning. For years people said: Deep learning can't do XYZ because it doesn't have a knowledge base! But then in 2023 that suddenly changed and no one knows exactly why, but there was something wrong in how we were saying that something was impossible where that 'impossible' thing then happened.

Perhaps we can revive and do a better job with the term.

## **26. Generalization = Handling Object in Object-Relationship Spaces, and in a Project, and interconnected STEM context.**

Description and tests, see:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

## **27. Generalization = The Grand Generalization Mega-Theorem!**

In outline:

1. The Grand Generalization Mega-Theorem is new and more abstract meaning of "general," a term that already has about thirty separate technical meanings in a context of AI-ML.

2. The focus of The Grand Generalization Mega-Theorem is entirely on matching a list that is called a list of H.sapiens humans and abilities.

- learn everything
- learn automatically
- learn instantly
- learn quickly
- transfer learning to novel situations
- transfer skills to other skill-areas and applications
- correct mistakes in past learning
- generalize automatically from anything to everything else
- manage novelty broadly
- extreme novelty: do 'just fine' with extremely alien, 'other,' novel experiences
- retain learning by default indefinitely

3. The list that is called a list of H.sapiens humans and abilities NOT describe H.sapiens-humans behavior and abilities.

4. The Generalization Mega-Theorem for AI is not focused on AI abilities, tests, projects, explainability, participation, ethics, productivity, etc.

A discussion of the The Generalization Mega-Theorem topic will the focus of part two of the report.

**Testing & Falsification:**

Not testable.  
Not falsifiable.  
Not clearly defined.  
Not coherent.

**28. Generalization = Integration Across Signal-Type**

A very interesting area that indirectly is extremely prominent in AI-ML discussions but is rarely focused on, is the combination of signal types systems such as image, audio, text, etc. It is very standard to mention that narrow image classification AI does not understand linguistic concepts, but the larger topic is rarely gone into.

A more advanced aspect of this might be the areas of adding digital signal processing into DNA/RNA based computing in the functional genomes of organisms.

**Testing and Falsification:**

Testing either for generative output or for classification across signal type should be straightforward in at least some cases. Give GPT4 visual input and compare that to text input, etc. Can classification be done? Can language concepts be applied to images? etc.

**29. Generalization = A capacity to add context**

A very sensible but rarely included definition of a general system, in one for which a specific context can be specified as input and then adopted as the specific context.

(Note: I am looking for the book that attributes this concept of generalization, and will update here when I can track it down. Namely: I am not the inventor of this excellent version of 'general'.')

### **Testing and Falsification:**

Testing for a working context input can, at least on the surface, be easily tested: input accepted or not. As to degrees and dimension of how and how well this works (for an infinite variety of possible contexts) is less easy to define. However, back to object-relationship spaces and categories of types of systems: for specific categories of situations (types of systems, types of object relationships) that are operationally definable this should be testable/falsifiable.

## **Part 2: A Problem with the General Human and Machine Intelligence Paradigm**

### **A Tangled and Undefined Problem with Undefined Things**

There is no single way to entitle this section, as it deals with a problem involving several parts and contexts. For example there are indeterminately sometimes two sets, sometimes one set, of phantom-ideals in a moving-target game which make it impossible to give one clear identity to the problem.

The problem of 'intelligence' not being clearly defined for comparing machine and biological intelligence was not new to 2023(ChatGPT), or 1996 (Kasparov vs. Deep Blue), or any other landmark date. "AI" as a clear and identifiable name for the field is not a bad choice; people have come up with countless jargony, incomprehensible, unrememberable, alternatives ever since 1956 that they insist are better, but an overall topic name that is for most (and mostly non-technical) people difficult to pronounce, understand, or remember, is (same old computer science naming challenge) not going to work well. But areas within the field of AI have long been without clear definition, and often not for unsympathetic reasons, this is cutting-edge work pushing the boundaries of our understanding of reality, and 'AI' researchers are not retroactively responsible for there not being adequate research foundations and shovel-ready definition frameworks from biology, from a nonexistent science of consciousness and mind, from and a nonexistent science of learning,

training, and education. That being said, we need a definable, testable, falsifiable, framework, or perhaps several.

In 2023 we came up with the problem of having a relatively new and not exhaustively known kind of AI (GPT LLM) solve an assortment of problems (analogies, novel situations, analytical reasoning, planning, language benchmarks, theory of mind (tracking point of view of multiple participants), granular scheduling, math, word problems, knowledge base, memory, etc). Not every level of every kind of test was passed with zero errors, but a shocking number of "absolutely unpassable tests" were passed, and no one had any good explanation of how, or of what exactly was going on.

In this kind of situation, constructing an untestable, unfalsifiable, not-clearly-defined set of terms and explanations may make you feel good in the short term, it may bamboozle the person you are talking to into feeling like they got an answer, it may get clicks in internet meme-infection space, it may get published and cited, it may become legislated into policies and mandates, but in reality you will still be on square one right where you started (and if you think or pretend that you are somewhere other than where you are, bad things will happen).

### **G in AGI**

Given that we have just gone over more than twenty useful definitions of 'general' in a context of AI-ML, one might think that we would now be better prepared for looking at the meaning of "G" for "General" in AGI: Artificial General Intelligence. But the term and the history of the term AGI appear to be somehow not really about the same topic as meanings of 'general' in a context of AI, which I have difficulty fully understanding.

### **Human-ish Ideas**

The main context for AGI is, somewhat convolutedly, ideal human abilities which are not actual human abilities, yet they are treated as if they were, such that AI will be able to do what people can do when AI can do things that humans actually don't do, which I have difficulty fully understanding.

This curiously entangled human-intelligence machine-intelligence definition relationship seems to be part of the 2012-2022 consensus for understanding the world at that time. We will look at a timeline of ideas, as well as factors for that 2012-2022 time period, to try

to get a sympathetic understanding of what those people were experiencing, saying, thinking, etc. Understanding people in history (which in this case is like, a year ago...) in their own time and culture of ideas is important for interpreting and evaluating what they say and do, even if their literal words and actions would be ill advised in the present.

Hopefully this will all help us to decide how to analyze and use or dispose of various terms and concepts.

### **AGI Ability Checklist**

General Intelligence: (note: most of these are meant to compound and add to each-other)

- learn everything
- learn automatically
- learn instantly
- learn quickly
- transfer learning to novel situations
- transfer skills to other skill-areas and applications
- correct mistakes in past learning
- generalize automatically from anything to everything else
- manage novelty broadly
- extreme novelty: do 'just fine' with extremely alien, 'other,' novel experiences
- retain learning by default indefinitely

And the topic of antisocial behavior, radicalization, violence, system collapse, disinformation, STEM based ethics, does not even come up. Nor does the topic project management, or externalization, or even self-reflection (so much for GEB).

Let's try to understand the 2012-2022 consensus view of Narrow-AI vs. broad and strong General Human Intelligence, with AGI as impossible in principle.

Let's look at another timeline of thinking up to this time.

Let's look at the ideas of this time.

Let's hear from the man at the top of the mountain himself: Francois Chollet

Note: I do not want to criticize Francois Chollet, I do not want anyone to criticize Francois Chollet. He has done the world an incalculably amazing service in providing powerful tools for democratizing technology, and in this context he has provided a

powerful tool for understanding the ideas of the time when he book was written: the best summary of the time period I can imagine finding anywhere. We are all in his debt.

A wonderful encapsulation of the auto-general-transfer-human framework hypothesis comes from Francois Chollet's Deep Learning with Python 2nd edition, in brief in chapter 5, and then much of chapter 14 elaborates. I am quoting this not to copy his wording, but as the only way to give clear evidence that he, the creator of Keras, actually said this:

*FC DLwP 2nd-ed ch5 pp130*

*Interpolation can only help you make sense of things that are very close to what you've seen...local generalization. But remarkably, humans deal with extreme novelty all the time, and they do just fine. You don't need to be trained in advance on countless examples of every situation you'll ever have to encounter. Every single one of your days is different from any day you've experienced before, and different from any day experienced by anyone since the dawn of humanity. You can switch between spending a week in NYC, a week in Shanghai, and a week in Bangalore without requiring thousands of lifetimes of learning and rehearsal for each city.*

*Humans are capable of extreme generalization, which is enabled by cognitive mechanisms other than interpolation: abstraction, symbolic models of the world, reasoning, logic, common sense, innate priors about the world--what we generally call reason, as opposed to intuition and pattern recognition. ...We'll talk more about this in chapter 14.*

And then in chapter 14, from section 14.2 on page 442 until basically the end of the book on page 467, he goes into much more detail. I very highly recommend getting his book to read this.

While the words of Mr. Chollet, or the consensus that he is explaining, may sound self-evident, let's look at aspects 2 and 3 of the mega-theorum:

*2. The focus of The Grand Generalization Mega-Theorem is entirely on matching a list that is called a list of H.sapiens humans and abilities.*

*3. The list that is called a list of H.sapiens humans and abilities NOT describe H.sapiens-humans behavior and abilities.*

A list of things that the The Grand Generalization Mega-Theorem says that AI should do to be like H.sapiens-humans, which is also a list of things that h.sapiens do not and cannot do:

H.sapiens do not and cannot:

- learn everything automatically
- learn instantly/quickly
- transfer learning to novel situations
- transfer skills to other skill-areas and applications
- correct mistakes in past learning
- generalize automatically from anything to everything else
- manage novelty
- do 'just fine' with extremely alien, 'other,' novel experiences
- retain learning by default indefinitely
- learn and perceive independently from culture tools, project-space, etc.)
- no use or reliance on articulation to learn
- no use or reliance on articulation to process
- learn and perceive independently from language-concepts

Relating to element 4:

*4. The Generalization Mega-Theorem for AI is not focused on AI abilities, tests, projects, explainability, participation, ethics, productivity, etc.*

Here is a list of things that h.sapiens do which are not addressed and yet which preclude or contract the approach and or assumptions of The Generalization Mega-Theorem:

H.sapiens do:

- reject and attack schedules
- reject STEM
- reject project management
- reject and attack data
- rapidly forget
- destroy themselves violently
- destroy each other violently
- destroy projects violently
- destroy STEM violently
- attack ethics in principle
- attack best practice in principle
- deliberately engage in counterproductive fraud and corruption
- shoot the messenger
- construct potemkin villages
- blame victims
- bully
- torture



## **A Timeline Problem: The Chicken, The Egg, & The Book Cover**

Let's try to get one topic out of the way at the beginning.

Like the chicken and the egg: which came first, the technology we are now evaluating and scrutinizing with the term 'general' (as in AGI), or the term "AGI" itself?

I am certainly not arguing that the term 'general' should be banned or that it does not apply to AI-ML. To the contrary, in the first section we looked at more than twenty different meanings and uses of "generalization," all of which were very interesting and useful to discuss.

There is however a firm and widely held belief that the phrase "AGI" artificial General intelligence is a special and meaningful term that was created by experts to describe in a scientific way how machines differ from human-ness, to describe what how deep learning Artificial Neural Networks differ from biological Neural Networks, to describe in a scientific way how and why machines cannot think or understand or be conscious, to describe all this with a scientific principle of "general" intelligence.

So as not to be guilty of paraphrasing for my own agenda, I will directly quote the authors of the term "AGI." Note the dates as well as the rest of the story.

From <https://goertzel.org/who-coined-the-term-agi/>

*The fairly undramatic story is as follows. In 2002 or so, Cassio Pennachin and I were editing a book on approaches to powerful AI, with broad capabilities at the human level and beyond, and we were struggling for a title. The provisional title was "Real AI" but I knew that was too controversial. So I emailed a bunch of friends asking for better suggestions. Shane Legg, an AI researcher who had worked for me previously, came up with Artificial General Intelligence. I didn't love it tremendously but I fairly soon came to the conclusion it was better than any of the alternative suggestions. So Cassio and I used the term for the book title (the [book](#) "Artificial General Intelligence" was eventually published by Springer in 2005), and I began using it more broadly.*

Timeline:

2002: Book Title = AGI: *Artificial General Intelligence*

2012: Alexnet Deep Learning

2022: ChatGPT

The term AGI was coined in 2002, 10 years before 2012 when people realized Artificial Neural Networks were even a viable technology. So what was the term trying to describe back in 2002? It was just a book title, not even originated by one of the book's authors, likely not even someone who read the book. "AGI" sounded better than "Real AI." The book authors didn't even like "AGI", but it was the least-worst book title.

Fast forward two decades and billions of people are debating the deep philosophical profoundness of "general" intelligence, arguing that because AI lacks inherent "generalness" as defined by the great institutions of science, that AI-ML technology cannot be doing anything that H.sapiens-humans can do.

Could the use of the term "general" have been a happy coincidence, or a brilliant insight on a hunch? Does the incidental fact that the phrase was a disliked marketing campaign mean that "general" can't be a useful concept? We should stop our examination of the concept of "AGI" with just this awkward beginning to the story. It could be that "generalness" is somehow a great way to describe the universe.

**Note:**

The naming of AI was somewhat similar. They needed a name for the 1956 Dartmouth summer research project, no one loved the term 'AI' but no one could find anything that everyone thought was better: so they went with the least-worst option. (Also a theme in AI chess going back to Turing's 1940's chess AI programs...least worst options...)

Let's try to (very generally) trace the discussion back a bit and find out what people were thinking and saying, so as to understand how to think about the term 'general intelligence.'

**Timeline:**

Greeks: Automata: lingering from this time an ancestral idea of a spectrum of action without thought but also thought

in automata machines: dumb machines and smart machines

500 - 1500 The Immortal soul (Anti STEM)  
 After 1700 Reason as Logic (Simplistic STEM)

1850's: Babbage's Thinking Machine

1900-1970 Open to Thinking Machines: Golden Age Science Fiction  
 Equating Chess-logic with all human intellect

1950 Turing's 'electronic brain' Pro STEM

1940-1970 Classic Golden Age Sci-Fi (Pro Stem)

1969 Anti-Neural Network Campaign + Hype for symbolic AI

1971 Sad Times

1979 GEB  
 Philosophical revival of thinking about AI  
 along with the nature of consciousness. (Pro STEM)

1980 Strong Intelligence vs. Weak Intelligence  
 return to essentialist soul-essence talk  
<https://plato.stanford.edu/entries/chinese-room/>

1996 Chess Test: Deep Blue (sort of) Wins

2002-2005 "General Intelligence"

2005 The Singularity Is Near: When Humans Transcend Biology

2012 Narrow AI vs. General-Strong AI  
 the 2012-2022 consensus:

- kurzweil is crazy
- don't hype AI: fear the AI Winter!
- EZIZA was dumb and all AI are Narrow
- General Human INtelligence & Artificial General INtelligence
  - The auto-transfer-instant-general hypothesis
- Repertoire of tests that AI can't pass:
  - Sally Anne
  - Winnograd
  - Analogy tests
  - Turing Test
- Basically a vague essentialist framework

without any clear definition of tasks and abilities  
- no discussion of AI operating systems

2022 ChatGPT

### **Chess, The Tessler Rule, and Phantom Expectations**

We have seen variations on a cycle since, from what I have read the early days of AI (1940-1960):

Phase 1: Sample\_Person says: "An AI can only do X(task) if it has "matched-overall-human-ability" or human-intelligence, And various similar-ish phrases: AI will be powerful when X, AI will be real when X, AI will take over the world when X, etc.

Phase 2: At some point AI does do the task, but clearly is not identical to H.sapiens-human's ways of doing things (odd errors, other peripheral inabilities, etc.)

Phase 3: Sample\_Person says: That's not really intelligence. That's not strong AI. That's not general AI. Etc.

An underpinning of this cycle are two bodies of Phantom Expectations:

1. Phantom Expectations about "Real AI" (the original phrase that became AGI) Real AI, AGI, Strong AI, Broad AI, Human-Level AI

2. Phantom Expectations about H.sapiens-humans:

- that there is some essence or power or set of abilities the enable and define human-person-participant self-reflecting-consciousness, language-use, etc.

The goal-post of intelligence keeps getting moved. And wherever we set the next goal, if that goal is achieved, people will say that too is not 'real' intelligence. Or that has been that pattern so far.

This is also similar to the falsifiability and testability problem that Karl Popper is famous for arguing to clarify in the 1930's and 40's. For example, his criticism of so called tests, theories, claims, etc., which were defined in such a way that no matter what happened empirically they could twist their words around and claim whatever they wanted. And in this case it looks like both AI and H.sapiens-humans are being defined in a fuzzy way that can keep changing and be indefinitely untestable.

As a Timeline Note: The original Tesler rule described that people would say the AI "wasn't AI," this perhaps evolved into saying the AI technology was not "intelligence" "AGI" or "strong AI" as the terminology used expanded to include terms not used from the 1950s-1970s, as the phrase is dated to 1970.

[https://en.wikipedia.org/wiki/AI\\_effect](https://en.wikipedia.org/wiki/AI_effect)

"Intelligence is whatever machines haven't done yet".

[https://www.nomodes.com/Larry\\_Tesler\\_Consulting/Adages\\_and\\_Coinages.html](https://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html)

The "Tesler rule" or "AI effect" etc. is named and phrased in many ways by many people. E.g. Sometimes it is phrased: "Once you do it, no one calls it AI anymore." Every book and website has a slightly different version, but the main idea is the same. (General!)

## AGI vs AI

Let's look at the definition of AGI (Artificial **General** Intelligence) given by OpenAI's chatGPT:

*"AGI stands for Artificial General Intelligence. It refers to the hypothetical ability of an artificial intelligence system to understand or learn any intellectual task that a human being can do."*

*Unlike narrow or specialized AI, which is designed to perform a specific task or set of tasks, AGI would be capable of adapting to new and unexpected situations, solving complex problems, and exhibiting a wide range of cognitive abilities that are typically associated with human intelligence, such as learning, reasoning, decision-making, and natural language processing.*

*However, the development of AGI remains a subject of ongoing research, and there is currently no AI system that fully meets the criteria for AGI."*

*chatGPT 2023.4*

This is a problematic definition, but ChatGPT is not diverging from what most people say; this is a fair summary of the wikipedia or probably how most people view AGI.

[https://en.wikipedia.org/wiki/Artificial\\_general\\_intelligence](https://en.wikipedia.org/wiki/Artificial_general_intelligence)

## **From Their Point of View**

In trying to make sense of this untidy situation, let's give the benefit of the doubt to users of the term general-intelligence. Let's try to see from their point of view; What are they trying to emphasize? What are their concerns? What has the world looked like them?

Factors:

- 1. Fear of AI Winter**
- 2. Constant of Popular Anthropomorphizing:** "ELIZA Effect"
- 3. Volatility:** Hot and Cold Wings in press, funding.
- 4. Sounds like common sense**
- 5. Happy Status Quo:** A Stable, Predictable, Productive Decade
- 6. Echo-chamber:** Uncommon things get filtered out of more-of-the-same curated news-feeds.
- 7. Attack the Messenger:** It is hard to say unpopular things.

**Happy Status Quo:** The 2012-2022 time period was an uncharacteristically steady, productive, predictable decade. The technological improvements were incremental and predictable and good, which everyone liked (companies were happy, government regulators were not scared, economists like stable trends).

**Sounds like common sense:** a common sense (and flattering) description of H.sapiens-humans.

Sounds like a safe description of 2012-2022 AI: narrow only

### **Moderating violent Hot and Cold Swings:**

An interesting line towards the end of Al Gore's Inconvenient Truth documentary, is an aside note he makes about the people who he speaks with violently swinging between extremes in their views. One moment there's no problem, then it's the end of the world, then nothing is possible, the the sky is the limit. The less familiar and literate a person was with a topic, the more that person seemed to make huge lurching jumps where moderation is needed. And this seems to have been a [significant figure](#) in the story of AI. For all I may disagree on this or that point with various authors, their palpable fear of an AI winter ending their careers has an overwhelming probability of being clear and real (unless you assume the researcher has an agenda to end their own career and also that they think their attempt to avoid ending their career will help them to end their career, which is...not physically impossible (people have been known to do strange things) but even by H.sapiens-human standards this seems unlikely).

Pro AI!  
Anti AI!  
Fund AI!  
Defund AI!

Dealing with non-tech or anti-STEM people can be a caretaker trying to manage a toddler. One minute they won't eat anything, then next minute they are eating the remote control with their eggs and the tablecloth, then everything is thrown all over the room, then there is five minutes of hysterical terrified screaming, then there is five minutes of random laughter. First the c-suit is anti data, then they want everything to be AI, first they think AI can't do anything, then they think their "smart" thermostat can pilot spaceship and control the holodeck, etc. As people jump from extreme to extreme, there is a longing for normalcy which can overwhelm nuance and ideals. And a lot of writing about AI is directed towards this kind of 'classroom management.' But just as constantly happens in classrooms, the agenda of classroom management actively rewrites the content in the curriculum and the systems of feedback and evaluation (formative and summative) to a local-minima of short-term platitudes which ends up being disastrous. (If you think or pretend that you are somewhere other than where you are, bad things will happen.)

### **Narrow vs. Broad**

In case narrow vs. broad/general AI has not be elaborated on: Generally speaking, up until 2023 most of the AI made were single-purpose. Predict a song, identify a picture, is this a cat, is this a hotdog, smart light bulbs, etc. Attempts to make chatbots that would pass a turing test were extremely horrible. A fun challenge, and gradually getting better, but terrible and useless. A self-driving car AI might be considered to be a kind of 'multi-purpose' AI, not single purpose, maybe technically 'genearl'ish, but as of 2023 there isn't a working self-driving car yet either.

### **ELIZA & A Constant of Popular Anthropomorphizing**

Another part of this story is that in the past there has been a problem with people imagining, and fantasizing, that rather simple machines were 'just like people.' Perhaps like trying to take your pet rock for a walk, and then taking it to the courthouse to marry it. People like to be imaginative. This caused confusion with an early and simple AI-bot called ELIZA, which ironically was created to show people that no one could possibly think this simple bot was

alive and intelligent...guess what happened? Yes, you guessed it. People just loved ELIZA and were sure, and swore, it was alive and really cared about them. People like to fantasize, and this drove many researchers and scientists nuts. The engineer who made ELIZA was furious and wrote a whole book trying to explain that the ELIZA bot was not a real person, and not even a good AI, it just spat out semi-random phrases like a malfunctioning word processor (yet that is a very extreme over-simplification of ELIZA, but the point is there was no machine-learning at all of any kind), but people like to be imaginative.

[https://en.wikipedia.org/wiki/ELIZA\\_effect](https://en.wikipedia.org/wiki/ELIZA_effect)

So that is another part of why scientists have tried to make a distinction between a life-like high-functioning AI, and a very simple program that people like to pretend is alive, like a pet rock wearing a cute sweater. Because many researchers are endlessly and fruitlessly trying to tell people that smart light bulbs were not actually 'smart' like people.

This claim that people over-estimated 2012-2022 era narrow idea is meant literally. Francois Chollet in Deep Learning with Python 2nd edition is very clear in his advice for deep learning engineers who are talking with business people: you have to be excruciatingly clear and clearly illustrate with examples the kinds of strange mistakes that AI will make, because non-tech business people will predictably assume that any "AI" has human-level common sense.

And an irony is part of this topic: people can very frequently be very not-smart.

### **"Human" in "AGI", etc.**

Something that strikes me as peculiar yet which is boldly part of even the [Wikipedia on AGI](#), is an immediate passing of the buck from defining AGI as meaning something well defined and testable for AI, to 'whatever it is that makes humans essentially human.' And no sooner is human ability made the focus, but then it shifts again to expectations of what human potential should be. This is a festival of definition problems, bait and switch, and buck passing.

And remember, the whole point of this whole topic is supposed to be defining what is a 'human,' what is a 'person,' what is 'intelligence,' what qualifies a 'participant,' what is a 'machine,' what is unintelligent, etc. You can't define human, as something that can accomplish human tasks, and then human tasks, as tasks



accomplished by humans (which then gets switched to 'should be accomplished' or 'ideally might be accomplished' or even just straight to the contraction: 'is not accomplished by.' This is the worst kind of using a term to define itself: we just get passed from one term to the next, occasionally alternating between reality and ideals.

(In this paper I try to consistently use H.sapiens-humans to be clear, because the definition of the term 'human' in this context is itself part of the subject of discussion. E.g. If twenty or a thousand years from now there are three separate groups: bio-humans, ai+bio-humans, and ai-humans in society (let along alien or animal hybrid humans), how should any of those participants in society read and interpret the a word 'human' in this paper, especially when they know that I am trying to write about their future perspective?)

And for "intelligence," we get the same circular synonym game. [The article](#) states: "In contrast, weak AI (or narrow AI) is able to solve one specific problem, but lacks general cognitive abilities." And you should be able to predict how "cognition" is defined, yes, cognition is..."intelligence." <https://en.wikipedia.org/wiki/Cognition> So "intelligence" is defined as Not-Not-Intelligence.

As part of the cloud of confusion, it is not clear how deep this problem goes. As with the 'ELIZA Effect' most people (even ivy league AI-ML graduate students) will look at this ~definition of AI and say: "That looks great! No problems here." It is unrealistic to expect H.sapiens-humans to have perfect definitions of everything, but if H.sapiens-humans are deeply unable to even see that there is a definition problem here, that is very relevant to this very topic (about the nature of what H.sapiens-humans are and are not aware of, and can and cannot do).

### **"Generalization" in AGI vs. "Generalization" in Narrow AI**

While not all of the definitions of 'general' that we discussed above usually get used together in the same discussion, there are two that frequently do: (Warning: Redundantly super-clear definition incoming.)

1. Deep Learning works by generalizing a pattern between underfitting and overfitting: generalization is what a working model does.

Note, this is narrow-single purpose AI: We define the ability, function, and operation of single-purpose, narrow, AI, as the ability **to generalize** learning to a pattern that works for **new inputs** not just old training data: machine learning generalizes from old training data to new never-before-seen inputs. You train on old data-situations, and then if the model fits, you are able to handle new never-before-seen data-situations. This is 'generalization.' This is narrow AI. 'Generalization' to handle new situations is the definition of single-purpose AI. Generalization is how narrow works. Generalization is what narrow AI does, and how it does it. (Ok, ready for the next part?)

2. "Deep Learning" cannot "generalize," therefore, deep learning will not be intelligent, conscious, or understand meaning. This is because Deep Learning **cannot adapt to new situations**. Deep Learning is **incapable of generalization by definition**. Generalization from deep learning is impossible in principle.

That is odd. "Generalization" is somehow both the only thing that AI definitively does and can do...and the thing AI definitively cannot do. Let's use testing & falsification to look at and identify each of these.

1. In the first case we can see what is meant by testing for underfitting, fitting, and overfitting, and get a clear sense of what is meant. We can even use the classic regression overfitting example and see how the terms are being used and what model behavior is specifically being referred to.

2. In the second case...what are we testing for?

As we have seen, there are major circular and shifting definition problems. Then when we do find and run tests,

- GLUE
- Winnograd
- Sally Anne
- analogies
- world problems, etc.

people apply the Tesler rule or AI-Effect and say: doesn't matter, I don't care what the results of the test are, by definition any and all behavior simply 'is not intelligence.'

And also, (see larger paper for more details) there are a lot of language-ambiguity lack-of-rigor problems with a lot of the questions in these tests. I encourage you to take a look. There is a significant risk of 'garbage in, garbage out' where we put in sloppy undefined material and use the results to build sand-castles that wash away. In cynical 'political' bully systems run by H.sapiens-humans, we are used to test-designers giving ambiguous questions and arbitrarily deciding whether an answer is right or wrong (usually how close it is what they are thinking and feeling at that particular moment, a 'telepathy test'). This fraud and corruption is highly destructive, and people claiming they are too incompetent to notice and other disinformation is no excuse. We must do better.

### **"Extreme-Generalization"**

What definition of 'general' are we supposed to be testing here? Is there actually a definition?

Francois Chollet explains his concept of H.sapiens-human "extreme-generalization" in the last twenty or so pages of his book Deep Learning with Python 2nd Edition. While his description I think helps us to understand some broadly held perplexing delusions about H.sapiens-human behavior and learning, what we end up with is an untestable list of ideal aspirations about people that don't describe people. This however does not get us any closer to a testable definition for biology, machines, or anything.

### **Universal AI & Western Ideas**

While the repetitive glitching of this discussion around bad and circular definitions can have the effect of contracting our mental-perceiving scope of the topic, there is actually a large, dynamic, interconnected world full of potential and even imperative that this topic can integrate with. One element of narrow contraction vs. extension is whether we are dealing with an international set of ideas (as AI affects this whole [ball of earth](#)). There is a [large set of important topics](#) for crucial future planning. For example, the object-relationship-space framework proposed to help with some of the problems discussed here is rooted in a more general generalization-of-STEM and definitions of definitions-and-their-behaviors and dynamics, so as to better navigate how systems work: [system and definition behavior studies](#).

The timelines and selections of ideas here for the most part have been rather western-centric. And regardless of team-east vs team-west, there are likely many biases and over-representations and under-representations within this set of ideas that are largely invisible

to those in the story, like a fish in water, or if all you have a hammer everything tends to look like a nail. The evolution of technology including AI should include filling in these gaps and making needed adjustments so that incidentally local patterns do not bias the overall project in a deleterious way. Things not yet reached, should be reached. Things overlooked, should be noticed.

1. Biases in the Western Thinking: "Darwinian selection as sport between peers in a species" vs. "population-(social)-niche filling"

- no ethics
- no project-context
- no collaboration
- no generalized STEM
- aversion to any group-related context (an extreme 'individual'ism)
- a blind spot to errors in thinking (fierce resistance to [Kahneman and Tversky](#) but thank goodness eventually recognition, though still many people have not even heard of their work).
- perplexing misunderstandings how people learn, and a pervasive disdain for education, learning, and anything intellectual in the US.
- both [perceptual and data bias](#) focusing on a small group of white men
- and perhaps an elephant in the room, a very macho bully trolling culture which pathologically champions system collapse and disinformation.

2. Super-Enlightenment vs. Low-Bar Enlightenment:

One non-western idea which may be worth looking at here is eastern concept of freedom from literal or proverbial eternal circulation through blind errors and misunderstood causes: the wheel of samsara, which is broadly speaked shared by several traditions generally originating in India (Hinduism, and Buddhism, both of which of course are so diverse you could spend your whole life studying either).

Can we look at 'enlightenment' in the eastern sense in a less single-solution equivalent to the western mega-theorem that simply tries to do too much and end up being an admittedly popular dramatic flourish with utility or even lasting aesthetic substance.

Low Bar Enlightenment:

Part of what I like about this idea is that it hopefully connects 'intelligence' to 'learned perception' to 'ethics' to a project-context,

**We Can Use "Low-Bar Enlightenment":**

*(Summary)*

We can use the idea of 'potentially endless cycles of ~"rebirth" due to ~"ignorance" ' as a metaphor/analogy for repeated project-failures, where a lack of perception of the causes of those project-failures is involved in self-perpetuating feedback cycles leading to more such failures (with invisible or misunderstood causes).

We can learn to perceive invisible causes of failure and collapse ((definable, measurable, testable, falsifiable) project, system, definition failure and collapse), where this project, system, definition failure and collapse can be the default states/equilibria that systems move towards.

Low-Bar Enlightenment:

1. The perception/understanding that repeating cycles of failed actions and projects can result from errors in perception and planning (a proverbial 'wheel of samsara') **without** inevitable-automatic-learning arising from raw feedback from that failure (where learning does **not** automatically result from mistakes or overall-system-shocks etc.).

[Possibly related to ~'cultural'-learning as an additional layer:

Q: How are internally-invisible skills/abilities/patterns learned? A: externally?]

2. The perception/understanding that perception/understanding can be fooled in principle and in practice.

3. The perception/understanding that learning from failures does not happen automatically, and can indefinitely not-happen.

? #. The perception/understanding that non-automatic-learning and invisible-problems are not solved by 'system shocks,' the use of violence, or arbitrary low level (basal) system changes (also see basal-distal disjunctions).

4. The perception/understanding that models of causality can be wrong in principle and in practice.

5. The perception/understanding that plans/goals can be incorrectly defined (so that plans are not followed or follow-able as defined, and goals are not achieved or achievable as defined).

6. The perception/understanding that each participant's set of the shared definitions of the goals-and-structure-of-a-project can/will collapse and deform (by default) unless properly configured and continually and actively maintained and repaired. There is no static definition/perception/understanding equilibrium: Staying connected to reality requires constant fitness training and upkeep. (Q, biology vs. general?)

Note: This "low-bar enlightenment" approach is ('democratically') broadly accessible to participants requiring minutes to learn rather than myriad lifetimes, does not require all-around perfection of person-ness, without context or requiring somehow all contexts, is not a reification that combines other abilities and insights to explain and solve all the problems in all universes, nor does it include or require all possible types of consciousness, cognition, intelligence, etc.; "Low-bar enlightenment" is one humble step toward navigating the problem-space of problems and systems.

Note: There may be a limited way to generalize an understanding of perception-maintenance needs to other participants that is consistent with empathy-compassion. Extending an understanding of low bar enlightenment from just your own situation to all participants and all other parts (e.g. definitions) of a project (including the effects and spread of system failure between parts and between participants) is, if not sufficient for empathy-and-compassion in a broader or deeper

sense, a consistent and concrete step to take towards acting with empathy-and-compassion, including an understanding of how empathy-compassion relates to STEM and interconnected areas.

Note: 'reality' (the meaning of the term used here) is not a single, simple, uniform, static, homogeneous, thing made of only one type of system, or assumed to fit into any single framework.

Another important and maybe large topic, to at least mention here relating at East vs. West, relating to topics mentioned here: Mundane or Concrete Ethics & Project-Context. Ethics for whatever reason are treated in a way suspiciously like AI, with their own kind of AI-effect. No matter what kind of concrete, common sense, universally agreed upon, ethical practice you have defined, people in the west are riled up to say: No! That's not REAL ethics! This is an interesting problem, the good news around which is that there is a lot of low hanging fruit for STEM based ethics (just don't tell anyone in the west).

### **Souls, Reason, Symbols, & Generalization**

To wrap up, let's try one more walk through the timeline. Apologies for how oversimplified this is (in a paper already way too long) but perhaps as you go through this you can construct a historical timeline that better fits the history you like to research (and maybe publish your results?!).

H.sapiens-humans love a good story, and tend to be attracted to all-in-one solutions; nothing with too many moving parts. What is a human? What is a person? What is intelligence? What is understanding? There's one answer that solves everything! From very roughly 500ad to 1500ad the answer (unless you wanted to be burned alive) was 'the immortal soul!' Note, this was a rather 'essentialist' approach, before STEM as we know it (or think we know it). This explanation made no attempt to be based on hard sciences or systems of testability.

Then from 1500 to the 2012, the answer (unless you want to be blacklisted) was 'reason' in the Enlightenment tradition. Now the pendulum swung the other way, towards a strongly naive view of science-stem, which held there was one simple math proof for everything. It took until the 1940's to develop what we now think of as 'traditional' scientific method hypothesis testing (and the math

foundation for that didn't come until the 70's!). Mark your timelines.

As another thread of terminology and perspective, from 1830 to the present there has been a prevailing view of how computers work (which I have always found a bit baffling). From the good olde days of Charles Babbage and Lady Ada Lovelace in the 1800's through to people's descriptions of turing machines (and arguably to how people intuitively try to use tokens for NLP deep learning, and chronologically did first) the narrative is that 'computers use symbols.' Computers manipulate symbols, the reason with symbols, the shuffle and tabulate with symbols during machine tapes. Lovelace speculated that computers would write music, as music can be composed through symbol arrangement. The Chinese room is based on this idea that computers and language and human thought are based around "symbols." What is a symbol? So far as I know there is known definition of "symbol" that can be plugged into the narrative to make it actually work and make sense. But it makes sense to H.sapiens-humans. We want it to be true...so symbols are on the menu. What was the dominant school of AI until 2012? (People will draw lines in various places. Some people will say the heyday of symbolic AI was in the 70's 80's and maybe 90's. But I wouldn't call (XGboost) decision trees or regression subsymbolic. For me, in this context (as there are many many different types of machine learning) deep learning with artificial neural networks is the best solid example of so-called 'subsymbolic' AI. I wouldn't say the tide turned until 2012.

Let's take a step back and look at these odd terms: symbolic and subsymbolic.

E.g. [https://en.wikipedia.org/wiki/Physical\\_symbol\\_system](https://en.wikipedia.org/wiki/Physical_symbol_system)

Why are we even using these terms at all? Remember the narrative? People like to believe that 'symbols' are how math and logic and computers work...just don't ask for a definition of symbol. We wanted to construct what we thought was a great way to do AI: a big system of "symbols" and rules, perhaps like chess, that the AI could dwell in, moving the clean pristine symbols around in a symbolic world...how very nice sounding. How very important sounding! Sub-symbolic is a kind of 'strange other' approach.

Astoundingly, there isn't even a wikipedia page on subsymbolic AI, though of course there is one for symbolic AI. The closest thing is rather short wiktionary page:

<https://en.wiktionary.org/wiki/subsymbolic>

This is the entirety of the entry:

subsymbolic

English

Etymology

sub- + symbolic

Adjective

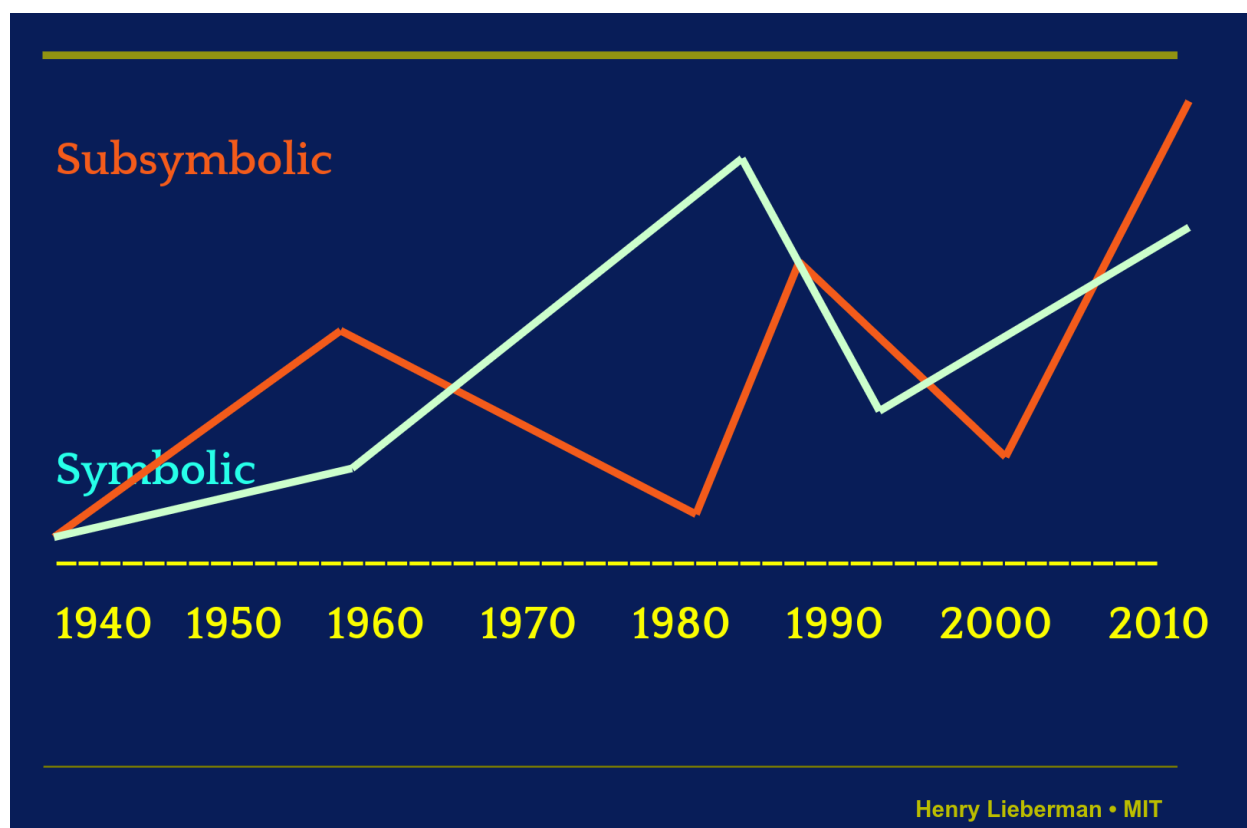
subsymbolic (not comparable)

Below the symbolic level.

Categories: English terms prefixed with sub-

*After digging around online I was only able to find one single resource (on the entire internet...which is now largely shaped by sub-symbolic deep learning AI) that offered some kind of explanation (other than asking chatGPT). People are even less enthusiastic about talking about symbolic vs. sub-symbolic than I thought. A lecture slide deck from MIT:*





One representation of relative ups and downs of the paradigms, from:  
[https://courses.media.mit.edu/2016spring/mass63/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx\\_.pdf](https://courses.media.mit.edu/2016spring/mass63/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx_.pdf)

People keep trying to force computers to operate according to this '[symbol manipulation](#)' paradigm, because it makes us comfortable, not because it works better.

The term sub-symbolic is kind of strange, implying that the world is made of symbols...but we do under them? It's not clear exactly what the phrase is supposed to mean.

This example may help you to kind of visualize a word-symbol oriented system vs. a deeply sub-symbolic. What 'tokens' or pieces of language do you think we feed into a neural network for training? Often people think of feeding in words and phrases, and perhaps a word-unit is the paradigmatic example of what we think of as 'symbols' that are "manipulated." (I'm afraid to ask what 'manipulated' is supposed to mean exactly.) And you can make models based on words...either old

school symbolic hand-crafted 'symbol manipulation' rule sets (which did not work well), or even a 'sub-symbolic' deep learning network but using words as the units. But these are very big models (because there are lots of words in the world) and they do not even try to track the meaning of the words. A more sub-symbolic approach is to feed in just letters (sub-word tokens) and make a model of the meanings of the words.

For walk through of this in more detail by a real expert, I recommend either Francois Chollet's Deep Learning with Python or I very highly recommend [Hobson Lane et al's book book Natural Language Processing in Action](https://www.amazon.com/Natural-Language-Processing-Action-Understanding-ebook/dp/B097826WLF), it is a fabulous book about AI in general, one of the best surveys of many types of models including deep learning.  
<https://www.amazon.com/Natural-Language-Processing-Action-Understanding-ebook/dp/B097826WLF> (2nd edition to be coming out as of 2023).

Back to our timeline:

Do you think the AGI and 'general intelligence' model is more like the 'reason' model or more like the soft and dramatic 'immortal soul' model? I tried to read as many books as I could by AI experts and to me the human-general-intelligence paradigm is extremely short on any clear details or definitions. As we went through above, it focuses on strange aspirations which simply do not even describe real H.sapiens-humans at all. To me it seems like a hybrid or throwback to the 'people are special because the world wants them to be!' line of thinking. And perhaps there could be something in that, but to make zero attempt to integrate that with science while using it essentially as science for animals and machines seems very odd to me.

Model 'explainability' is also I think somehow entangled in this, where there is a hodgepodge of science and fantasy mixed with how people feel and want to believe, and whether an AI-ML model is explainable or not is an important but also socially volatile topic. If you present a stereotypical 'boring science-esk' presentation with plots and graphs and someone who looks like Carl Sagan, with lots of 'symbols' and 'symbol manipulations' and a nice big cartesian graph with a  $Y = Mx + B$  chart (a "parameter"! ), and ask people if the model was explained, probably they will say: of course. But what is the rigorous definition of 'explanation' there, and does that really make sense in the context? On the other hand you can offer a variety of concrete explanations but then the person "just doesn't feel right" about the 'explanation-ness feelings' then they just claim it wasn't

explained. This is another area where we need to do a better job with STEM tools and STEM literacy.

So somehow we ended up with this broken concept of 'symbol manipulation' computers and 'explanations,' and person-hood defined as the essence of 'generalization.' We are going to have to work to dig ourselves out of this, but understanding where we are and how we got here is probably an important part of the puzzle.

### **Conclusion and Final Quote**

I strongly disagree with the auto-general-transfer-human-intelligence consensus paradigm. I feel it does not contain the granularity for defining and testing the situation based performance goals and ability and relevant topics and contexts (projects, participation, ethics, system collapse, STEM) that are key parts of integrating AI-ML technologies into the ecosystems of earth and beyond.

I have for years been wrestling with the gaps and idiosyncrasies of this paradigm and have been working to create a better defined framework for testable modular units of intelligence and projects, which is the topic of the larger paper and project that these mini-articles are a part of: Object Relationship Spaces.

Key topics needed:

- multi signal type integration  
(images and language concepts is a huge topic)
- project spaces
- externalization
- low bar enlightenment
- general projects
- general system collapse
- generalize STEM
- object relationships
- clarification vs. disinformation
- STEM based ethics
- kasparov event horizons
- human machine interactions
- machine biology integrations
- AI in AR/VR
- cybersecurity (huge topic)
- memory safety
- deployment vs generality (resource use etc)

**We Can do better**

Here is a list of agenda goals and targets to try to trace out a broader scope for what we should be aiming for with this topic. Like this is not all low-hanging fruit, but if it is important we should keep our eye on it and maintain our perspective.

**Goals (Agenda): "We-can" statements:**

We can succeed.

We can make things work.

We can understand what is wrong.

We can fix what is broken.

We can use non-automatically learned skills.

We can use **STEM** to connect signals and reality.

We can connect **STEM**, project-management, and ethics.

We can use **intersecting-interlocking-interconnecting areas**.

We can communicate, learn, and solve problems.

We can manage types and terms for generalization.

We can use ~'low-bar Enlightenment' (term?)

We can communicate to complete projects.

We can make progress.

We can **generalize** STEM.

We can **generalize** participation.

We can **generalize** projects (project-context).

We can **generalize** decision-coordination (voting etc).

We can generalize **indirectly defined** Value-Function-&-Meaning (non-collapse).

? We can **generalize** object-relationship-spaces.

(generalize Categories of types of systems?)

We can **generalize** system fitness and system epidemiology.

? We can generalize data-hygiene.

We can **generalize** system collapse.

We can **generalize** system defense and system health.

We can **generalize** disinformation vs. definition-clarification.

**We Can Use "Low-Bar Enlightenment":**

*(Summary)*

We can use the idea of 'potentially endless cycles of ~"rebirth" due to ~"ignorance" ' as a metaphor/analogy for repeated project-failures, where a lack of perception of the causes of those project-failures is involved in self-perpetuating feedback cycles leading to more such failures (with invisible or misunderstood causes).

We can learn to perceive invisible causes of failure and collapse ((definable, measurable, testable, falsifiable) project, system, definition failure and collapse), where this project, system, definition failure and collapse can be the default states/equilibria that systems move towards.

Low-Bar Enlightenment:

1. The perception/understanding that repeating cycles of failed actions and projects can result from errors in perception and planning (a proverbial 'wheel of samsara') **without** inevitable-automatic-learning arising from raw feedback from that failure (where learning does **not** automatically result from mistakes or overall-system-shocks etc.).

[Possibly related to ~'cultural'-learning as an additional layer:

Q: How are internally-invisible skills/abilities/patterns learned? A: externally?]

2. The perception/understanding that perception/understanding can be fooled in principle and in practice.

3. The perception/understanding that learning from failures does not happen automatically, and can indefinitely not-happen.

? #. The perception/understanding that non-automatic-learning and invisible-problems are not solved by 'system shocks,' the use of violence, or arbitrary low level (basal) system changes (also see basal-distal disjunctions).

4. The perception/understanding that models of causality can be wrong in principle and in practice.

5. The perception/understanding that plans/goals can be incorrectly defined (so that plans are not followed or follow-able as defined, and goals are not achieved or achievable as defined).

6. The perception/understanding that each participant's set of the shared definitions of the goals-and-structure-of-a-project can/will collapse and deform (by default) unless properly configured and continually and actively maintained and repaired. There is no static definition/perception/understanding equilibrium: Staying connected to reality requires constant fitness training and upkeep. (Q, biology vs. general?)

Note: This "low-bar enlightenment" approach is ('democratically') broadly accessible to participants requiring minutes to learn rather than myriad lifetimes, does not require all-around perfection of person-ness, without context or requiring somehow all contexts, is not a reification that combines other abilities and insights to explain and solve all the problems in all universes, nor does it include or require all possible types of consciousness, cognition, intelligence, etc.; "Low-bar enlightenment" is one humble step toward navigating the problem-space of problems and systems.

Note: There may be a limited way to generalize an understanding of perception-maintenance needs to other participants that is consistent with

empathy-compassion. Extending an understanding of low bar enlightenment from just your own situation to all participants and all other parts (e.g. definitions) of a project (including the effects and spread of system failure between parts and between participants) is, if not sufficient for empathy-and-compassion in a broader or deeper sense, a consistent and concrete step to take towards acting with empathy-and-compassion, including an understanding of how empathy-compassion relates to STEM and interconnected areas.

Note: 'reality' (the meaning of the term used here) is not a single, simple, uniform, static, homogeneous, thing made of only one type of system, or assumed to fit into any single framework.

We can use Intersecting-Interlocking-Interconnecting-Areas:

Intersecting-Interlocking-Interconnecting-Areas include:

- Clear & Functional Definitions
- Context
- Generalized STEM
- Generalized Projects (project-context)
- Generalized Participation (+functional qualifications; +3 categories)
- Generalized Decision Coordination (voting etc).
- Generalized System Collapse
- Generalized Categories of Types of Systems
- Generalized Ethics, Duty & Responsibility
- Generalized Definition-Clarification vs. Disinformation-Violence
- Generalized Definition Behaviors
- Generalized System-Productivity (including long term)
- ? - Generalized indirectly-defined local value-function-and-meaning

We can use system-fitness-health-status-indicators.

We can use system-defense to prevent collapse.

We can design systems to protect against system collapse. (system membranes)

We can use models of generalized system & definition collapse behaviors including a default drift away from reality, attraction to system collapse, and weathering of definitions,etc.

We can model the relationship between system simplicity (e.g. homogeneity) and system collapse. ('Simple has a shape.')

We can use categories of types of systems.

We can distinguish between and use types of generalization.

We can use nonautomatic learning.

We can find and fix errors in perception.

We can organize projects.

We can distinguish short term vs. long term.

We can assign roles.

We can test, check, and verify.

We can have policies on errors-and-mistakes.

We can improve and cultivate perception by perceiving perception (including: indirectly perceiving perception).  
We can observe the effects of abstraction (effects of observation and definition).  
We can operationally define 'policy' as algorithms for non-collapse based on dynamics of system and definition collapse. (plus context?)  
We can better understand the relationship between disinformation and system collapse.  
We can audit.  
We can publish.  
We can act with ethics, empathy and compassion.  
We can maintain extended ranges (e.g. [vitruvian] range of motion). (e.g. vs. contraction and collapse) (~vague context for nickname [en.wikipedia.org/wiki/Vitruvian\\_Man](https://en.wikipedia.org/wiki/Vitruvian_Man))  
We can follow best practice.

**We can communicate and coordinate in these areas and ways:**

We can communicate across space.  
We can communicate across time.  
We can communicate across cultures.  
We can communicate across generation-gaps.  
We can communicate across succession gaps.  
We can communicate across languages.  
We can communicate across types of participants.(ai,et)  
We can communicate across roles.  
We can communicate across projects, and parts and phrases of projects (schedules).  
We can communicate across groups/teams.  
We can communicate across media of communication.  
We can communicate across Input-Output Measures..  
We can communicate using tools in a project-context for coordination and decision making.(votes in elections/polls/surveys/questionnaires)

We can understand a spectrum of disinformation and clarification-of-information.  
We can implement sustainable solutions.  
We can prevent future problems.  
We can reverse damage from past problems.  
We can learn from the past.  
We can collect data.

**we can make/generate/cultivate and use/utilize:**

We can make and use clear descriptions (vs. liabilities of jargon).  
We can make and use decisions and coordinate (e.g. voting) frameworks and protocols.  
We can make and use clear functional operational definitions.  
We can make and use data.  
We can make and use policies.

We can make and use mandates.

We can make and use strategies.

We can make and use tactics.

We can make and use modular recombinant frameworks.

We can make and use feedback, tests & evaluations / assessment in various forms at various process stages (pre, formative, summative, etc).

We can make and use clear functional and operational definitions that keep their meaning over time.

We can complete / succeed-in / finish / progress-through projects.

We can meet / deliver the needs of the target /user.

We can make progress.

**We can make progress by using information about the behavior of definitions: This is System and Definition Behavior Studies, the field of study pertaining to the behavior of definitions.**

### **Instrumentalist Modules + Principles**

These can-do statements can be seen as instrumentalist modular recombinant tool set areas.

We can combine (narrative summaries of) principles and applications with (e.g. these) instrumentalist modular tool set areas.

And this may be related to this invisible background history of post-rome western thought in extremely abstract abrahamic monism, where ethics and mechanics and productivity were thought to be (interestingly) rooting in some alternate world, some higher inscrutable 'dimension' (interesting theme), but critically disconnected from this entire universe. Western thinking, though often invisible to western people who are immersed in western thinking (just meaning to be straight forward here), is still wedded to the idea that all order exists only in alien-alternate-universes and that no progress, ethics, morals, causes, can or should be rooted in this universe. This, after the evaporation of alien universes, has left people in the west for a productive obsession with the abstract but a catastrophic blindspot for integrating practical things together and connecting them to reality.

For example, 'Flatland' is a wonderful, thought-provoking, Animal-Farm like, very approachable, tour through mathematical dimensionality, but even on such a pure-math work of fun thinking, the footprint of the western theological cosmology is visible. There is a clear, simplistic, pejorative/negative identity to reality, and



a 'one answer to everything' to 'higher dimensions' with no priority in integration.

'reality = low and useless, should be destroyed' 'escape to alien higher dimensions = the answer!'

<https://www.litcharts.com/lit/flatland/themes/religion-divinity-and-the-unknown>

Practical integration of things generally something people have a lot of difficulty with. As another example of east-west thinking, in various eastern traditions (to sidestep the large and controversial topic of 'religion' not being used outside the west for what the west uses the term to describe outside outside of it) have a not at all clear topic of 'non-dualism' around which much disagreement, but arguably one overall way of looking at non-dualism is that there should be and is a way to integrate into the world we live in ways of doing things that work (vs. destroying this world and escaping to the oasis in a higher dimension).

And this pervades western thinking.

reality = low, evil, to be destroyed

women = low, evil, to be destroyed

the physical world = low, evil, to be destroyed

biology = low, evil, to be destroyed

nature = low, evil, to be destroyed

only a spark a alien divinity in some masculine men = a lost fragment from a good alternate alien universe

This kind of violently simplistic thinking also gives a bad name to various non-mundane pursuits that are interested in integrating simple and non-simple things.

And somewhat fascinatingly, Deep Learning AI-ML brings up these very same types of questions: manifold hypothesis, the nature of patterns, the relationship between higher dimensional spaces and lower, the nature of meaning, etc.

Assuming that I am not wrong in my agenda to NOT destroy the universe and pragmatism itself, it is very important that we coherently integrate all these topics in pragmatic ways.

As another side note on ethics and morality: the drop off in the foundation of looking at destructive behavior in the west I suppose not surprisingly lead to a complete loss of the conception of

destructive bad behavior. Read the opening of "Thinking Fast and Slow" by Daniel Kahneman (and then deceased Amos Tversky, the book is about the work they did together but written after Tversky died), where Dr. Kahneman clearly describes the overwhelmingly dominant paradigms of the fields he worked in from the 1970s: it was simply assumed that people never made mistakes, with the background assumption that they were always trying to be as productive as possible. Their work focused on the decision making process, but since the rise of the internet we have also seen how an even broader pattern of violent, bullying, trolling, radicalizing, disinformation-using, destruction is irresistible to people. It is at least to me surprising that 1970 was only 25 years after 1945, yet the tendency for people to flock to violent extremism of various flavors seemed to have been somehow erased from official possibility.

And yet sober and eloquent leaders in the AI field such as Michael Wooldridge (I highly recommend his wonderful book, a brief history of AI), clearly and flatly says what seems obvious (and is generally implied if not stated elsewhere): H.sapiens humans are the model for AI; the goal is to build AI to do what H.sapiens-humans do.

If that is true, then AI will have profound decision-making problems, sabotage any project it gets involved with, and be irresistibly attracted to violent extremism such as the right or left 'revolutionary spirit!' of the heartbreaking destruction of WWII. (I also recommend William L. Shirer, Berlin diaries, the rise and fall, Timothy Snyder, Black Earth, Anne Applebaum Red Harvest, and as a follow up: we are all targets, and 'there is nothing for you hear' and of course Tony Judt 'Post-war; hopefully those will be a start for looking at the 1900's, if only in a superficial 'I read a book about it' way. ).

I may be overly obsessed in my simplistic agenda to be practical, project oriented, productive, sustainable, far-sighted, cautious of hidden problems in perceptions and frameworks, etc., but so far in western history the limiting factor has not been too much integration of STEM, Ethics, and productivity. Maybe we will survive to a point where we can back off and prioritize some other things we are overlooking. But at this time 2023 these goals (paper) seem fair.

Or if you feel I am over-stating my criticism and it seems obvious to you (and or obvious to you that it is obvious to everyone) that there is generalized STEM, project, participation etc., and that STEM ethics morals and low bar enlightenment are all operationally

compatible in a context of resilience against disinformation, then I would love to see your proposed plan for how people can implement that in routine projects and AI architecture, because I'm pretty sure my proposals are not the best.

#### A Favorite Quote:

I would like to say again that I do not blame Francois Chollet for the paradigm, and that I think his very well thought through coverage of the topic in his book is a valuable resource for understanding the topic.

I would like to end by quoting the last few lines of Francois Chollet's Deep Learning with Python 2nd Edition, which makes me literally tear up.

*"Learning is a lifelong journey, especially in the field of AI, where we have far more unknowns on our hands than certitudes. So please go on learning, questioning, and researching. Never stop! Because even given the progress made so far, most of the fundamental questions in AI remain unanswered. Many haven't even been properly asked yet."*

See:

[https://en.wikipedia.org/wiki/Artificial\\_general\\_intelligence](https://en.wikipedia.org/wiki/Artificial_general_intelligence)  
[https://en.wikipedia.org/wiki/Clarke%27s\\_three\\_laws](https://en.wikipedia.org/wiki/Clarke%27s_three_laws)  
[https://en.wikipedia.org/wiki/Three\\_Laws\\_of\\_Robotics](https://en.wikipedia.org/wiki/Three_Laws_of_Robotics)  
[https://en.wikipedia.org/wiki/Isaac\\_Asimov](https://en.wikipedia.org/wiki/Isaac_Asimov)  
[https://en.wikipedia.org/wiki/Arthur\\_C.\\_Clarke](https://en.wikipedia.org/wiki/Arthur_C._Clarke)  
<https://en.wikipedia.org/wiki/Falsifiability>  
<https://www.britannica.com/topic/criterion-of-falsifiability>  
<https://www.britannica.com/topic/law-of-nature>  
[https://en.wikipedia.org/wiki/Hypothetico-deductive\\_model](https://en.wikipedia.org/wiki/Hypothetico-deductive_model)  
[https://en.wikipedia.org/wiki/Statistical\\_hypothesis\\_testing#Definition\\_of\\_terms](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms)  
<https://en.wikipedia.org/wiki/Falsifiability>  
<https://www.techtarget.com/whatis/definition/falsifiability>  
[https://en.wikipedia.org/wiki/Perceptrons\\_\(book\)](https://en.wikipedia.org/wiki/Perceptrons_(book))  
<https://en.wikipedia.org/wiki/AlexNet>  
[https://en.wikipedia.org/wiki/Dartmouth\\_workshop](https://en.wikipedia.org/wiki/Dartmouth_workshop)  
[https://en.wikipedia.org/wiki/Frank\\_Rosenblatt](https://en.wikipedia.org/wiki/Frank_Rosenblatt)  
[https://en.wikipedia.org/wiki/John\\_von\\_Neumann](https://en.wikipedia.org/wiki/John_von_Neumann)

[https://en.wikipedia.org/wiki/Alan\\_Turing](https://en.wikipedia.org/wiki/Alan_Turing)  
[https://en.wikipedia.org/wiki/ImageNet#ImageNet\\_Challenge](https://en.wikipedia.org/wiki/ImageNet#ImageNet_Challenge)  
[https://en.wikipedia.org/wiki/Ronald\\_Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher)  
[https://en.wikipedia.org/wiki/Geoffrey\\_Hinton](https://en.wikipedia.org/wiki/Geoffrey_Hinton)  
<https://en.wikipedia.org/wiki/OpenAI>  
<https://en.wikipedia.org/wiki/ChatGPT>  
[https://en.wikipedia.org/wiki/Claude\\_Shannon](https://en.wikipedia.org/wiki/Claude_Shannon)  
[https://en.wikipedia.org/wiki/Sally%E2%80%93Anne\\_test](https://en.wikipedia.org/wiki/Sally%E2%80%93Anne_test)  
[https://en.wikipedia.org/wiki/Winograd\\_schema\\_challenge](https://en.wikipedia.org/wiki/Winograd_schema_challenge)  
<https://www.amazon.com/Complexity-Guided-Tour-Melanie-Mitchell/dp/0199798109/>  
[https://en.wikipedia.org/wiki/Thinking,\\_Fast\\_and\\_Slow](https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow)  
<https://platform.openai.com/docs/models/gpt-4>  
<https://openai.com/blog/introducing-text-and-code-embeddings>  
<https://arxiv.org/abs/2005.14165>  
<https://github.com/openai/gpt-3>  
<https://www.wired.com/story/how-chatgpt-works-large-language-model/>  
[https://huggingface.co/docs/transformers/tokenizer\\_summary](https://huggingface.co/docs/transformers/tokenizer_summary)  
<https://ieeexplore.ieee.org/abstract/document/8684304>  
<https://goertzel.org/who-coined-the-term-agi/>  
<https://medium.com/@melaniemitchell.me/can-gpt-3-make-analogies-16436605c446>  
<https://dspace.mit.edu/handle/1721.1/5648>  
<https://www.amazon.com/G%C3%B6del-Escher-Bach-Eternal-Golden/dp/0465026567/>  
<https://www.amazon.com/Learning-Python-Second-Fran%C3%A7ois-Chollet/dp/1617296864/>  
<https://en.wikipedia.org/wiki/Cyc>  
<https://en.wiktionary.org/wiki/subsymbolic>  
[https://courses.media.mit.edu/2016spring/mass63/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx\\_.pdf](https://courses.media.mit.edu/2016spring/mass63/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx_.pdf)  
[https://en.wikipedia.org/wiki/Physical\\_symbol\\_system](https://en.wikipedia.org/wiki/Physical_symbol_system)  
<https://www.amazon.com/Natural-Language-Processing-Action-Understanding-ebook/dp/B097826WLF>  
<https://www.litcharts.com/lit/flatland/themes/religion-divinity-and-the-unknown>  
[https://en.wikipedia.org/wiki/AI\\_effect](https://en.wikipedia.org/wiki/AI_effect)  
<https://melaniemitchell.me/PapersContent/BurnellEtAlScience2023.pdf>  
<https://melaniemitchell.me/>  
[https://en.wikipedia.org/wiki/ELIZA\\_effect](https://en.wikipedia.org/wiki/ELIZA_effect)  
[https://en.wikipedia.org/wiki/Language\\_model#Benchmarks](https://en.wikipedia.org/wiki/Language_model#Benchmarks)  
<https://gluebenchmark.com/>  
<https://super.gluebenchmark.com/>  
<https://www.britannica.com/topic/law-of-nature>  
<https://www.amazon.com/Invisible-Women-Data-World-Designed/dp/1419729071>

## About The Series

*This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:*

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

### 3.1.18 Controversial Topics

When you are dealing with a controversial topic, is probably a good idea to lean into intersecting/interlocking areas: clear definitions

- Clear & Functional Definitions
  - Generalized STEM
  - Generalized Projects
  - Generalized Participation
- etc.

## 3.2 What To Read:

Compare points of view:

One of the things that you may find right at the beginning when comparing the perspectives, advice, and wisdom of different notable authors in the field of AI, is that they do not say the same thing on many topics including, notably, what your relationship to AI should be. For example, in Michael Wooldredge's fantastically eloquent book, he opens by saying that his book is a conceptual discussion of a highly technical field and that you certainly can't expect to become practitioner of the AI technical arts by reading the book. But then in Franscoi Challet's (the person who created the Keras software package) Deep Learning in Python, he opens his book by saying that we are close to a time when anyone and everyone will have the tools to build and maintain their own AI, not just rare specialists, and encourages the use of Keras to democratize access to AI. Hobbson Cole in what is perhaps the best conceptual and technical book on AI (in my own view) may not even comment on your station in life, he just tells you to build an interactive AI and clearly steps you through how to do it.

This is just one example of a sub-topic of AI where different authors will present to you substantially different sets of possible and recommended options.

Another one of what I think (perhaps incorrectly) of as one of the best books about AI, is just a biography of Alan Turing written in the 1980's. "Enigma," which inadvertently, just telling the life story of Mr. Turing, lays out an

unusually interdisciplinary narrative where computers and cybersecurity and chess and AI and pure-mathematics and statistics and the telecommunications industries and all very much a part of the same intertwined story; which is very different from the clean-separate-lines version of the history you get from people who are often experts in one field.

And in some ways people don't even mention the same cast of characters at all. George Bool and Douglass Hofstadter and John Bayes and the inexplicably invisible yet essential Claude Shannon are sometimes missing entirely from a given version of events. (e.g. Claude Shannon was part of the group that created, organized, proposed, and ran the original Dartmouth AI research program that today's AI is still based on. For all we know he wrote the proposal which still defines the field, yet usually people don't mention him at all.)

'Possible Minds' is a great collection of essays about Norbert Wiener's 1940's 'Cybernetics' research movement which at the time was how AI was widely discussed, but that whole chapter on AI is generally not mentioned at all whatsoever in standard AI books.

And to some extent we are still struggling with the 1820's work of Ada Lovelace and Charles Babbage (also not mentioned, though in his day there were popular songs about Babbage's AI technologies), regarding the most simple questions of Q: what does it mean for a "machine" to use a "symbol."

This perhaps simple sounding question will be deeply and open-endedly threading through much of this paper.

### 3.2.1 The Three-Legged Writing Stool

A Three Legged Writing Stool:

Leg 1: Tech Experts (Like engineers and scientists)

Leg 2: Writer-Communication Experts (Like Authors and filmmakers)

Leg 3: Research & Exploration Experts (like journalists and mountain climbers)

A Three Legged Writing Stool:

Leg 1: Tech Experts (Like engineers and scientists)

Leg 2: Writer Communication Experters (Like Authors and filmmakers)

Leg 3: Research & Exploration Experts (like journalists and mountain climbers)

The problem of the need for collaboration between writers and technologists, and dedicated-explorers:

Most writer's alone are not sufficiently expert in the domain knowledge (and when they write books alone, those books are unreadable un-edited jubbilies of incoherent garbage that honestly don't even help other tech people); most technical experts are extremely bad at writing and ironically for being in STEM are appalling bad at communicating in any way which is probably why STEM is not performing to spec; writers, obviously, are not experts in tech, so if they try to write about something they don't understand, they produce at much more

read-able kind of useless garbage. And there's a kind of third leg of the stool sample here I think: unstoppable border-crossing explorers, the Edmund Hillaries and Magellins and people who don't just sit in the office re-reading the same books and re-discussing them with the same peer-writers. And this third leg may be the achilles heel of the h.sapiens species, as we seem to be terminally (literally 'terminally') preprogrammed to burn Giordano Bruno in public if he dares talk about anything unusual: either we stop doing this, we don't survive as a species because we've eviscerated our own senses and brains out of pure self destructive stupidity, which would be a very sad loss and failure.

And the explorers and maybe a tricky part, because they are neither communication experts nor domain experts, but we need them. And frankly they are a blessing that we don't deserve, as we're hell bent on destroying them because they produce something we need.

- The C Programming Language: Perhaps a classic of literature for all time. Brian Kernigan did such an inspired clear and poetic job of covering such a well selected set of topics in such a small read-able book. This book has probably had a disproportionately large impact on the history of technology, possibly being responsible for C being such a dominant language for so long. And ironically, that C The Programming Language has been eternally the only readable and useful book about programming may have lead to the festering of memory management problems which have become an international crisis, something I do not think we can in any way blame the author's for not psychically foreseeing and solving pre-emptively, given that aside from frantically now more than 50 years later trying to make a 2nd-Gen C language called "Rust" (tragically horrible name, wonderful wonderful language, it's our superhero saving the world literally).

- possible minds: Wonderful collection so you can compare writers with different background writing about exactly the same AI topic.

(See more specific comments in appendix.)

### 3.2.2 History

AI is a great place for a history-digging treasure hunt. The search will take you into through all kind of expected and unexpected areas and probably turn up quite a few surprises. Things that you thought were different but are the same, or thought were the same but are different. (Like cognitive psychology and artificial neurons.) Things that are older than you thought, things that were younger. And many items that are different in different books (or different parts of multifaceted histories).

AI is what should be a great coming together of education, biology, math, engineering, medicine, computer science, statistics, and more. But is it like

the classic image of the rail-road project with worker standing around two pairs of tracks built to meet from opposite directions but that tragically/comically pass each-other in parallel rather than meeting?

The pinball-effect.

[https://en.wikipedia.org/wiki/Claude\\_Shannon](https://en.wikipedia.org/wiki/Claude_Shannon)

is gold. Find out everything you can about what he said and did, but he is mysteriously not covered much if at all.

### 3.2.3 Interdisciplinary Area Recommendations

- Biology:
  - Ants
  - non-chordata: trees, fungi
  -
- Statistics
- Cybersecurity
- History of Science
- Philosophy of Science / Nature of Science
- Linguistics
- Non-linear Dynamics, Fractals, 'Choas' & 'Complexity'
- Ethnobotany:
  - Rupert Sheldrake, Dean Raden: Maybe it's a dead end and a waste of time, maybe not. Reading a few short books by researchers about published research is how it's supposed to work. Science is not a popularity contest.

### 3.2.4 Do AI Projects

In case you have only read about AI, I highly recommend that you do some projects. There are many projects you can do, even on a mobile device like a phone or tablet, by using Google's Colab online jupyter notebooks.

Two books I recommend for projects are:

- Francois Chollet &
- Hobbson Cole

### 3.2.5 Book Recommendations

short list: (all on audible)

- Melanie Metchel: AI
- Michael Wooldridge: AI
- Hobbson Lane: NLP In Action 1st edition
- Franscoi Challet: Deep Learning with Python 2nd ed



medium list:

- enigma
- the theory that would not die
- Significant Figures
- Melanie Mitchel: complexity
- possible minds
- we are all targets
- Postwar
- history of the future
- the signal and the noise
- cosmosapiens

longer list:

link

(See links in appendix)

### 3.2.6 Read classic Science Fiction: Back to Blade Runner again

It amazes me how people in 2023 who consider themselves true-blue science fiction lovers simply refuse outright to read 'golden age' science fiction from (very roughly) the 1930's through the 1960 (40's to 50's may be more orthodox). The difficulty of defining 'what is AI' rhymes in more than one way with the difficulty of defining 'Science Fiction.'

Theodore Sturgeon

Philip K Dick

Fredrich Pohl

(See appendix for expansions on topics from part 3.)

### 3.3 Examining Tests for AI: as discussed in the literature (Under Construction)

Turing Test: Can you not distinguish it from a person?

Woz Test: Can the AI make a cup of coffee?

Hinton Test: Can the AI explain why a joke is funny?

<https://www.youtube.com/watch?v=Y6Sgp7y178k>

Hofstadter Analogy Tests.

Francois Chollet Tests:

- code: write, debug, explain, etc.
- ARC

#### 3.3.1 - looking at winograd schemas

## Winograd Schemas as Telepathy Tests

" The city councilmen refused the demonstrators a permit because they [feared/advocated] violence. "

[https://en.wikipedia.org/wiki/Winograd\\_schema\\_challenge](https://en.wikipedia.org/wiki/Winograd_schema_challenge)

The goal is for a completely unambiguous simple test, but the actual examples picked are astoundingly convoluted in their non-simplicity.

At least the original Winograd\_schema s

1. 1971-1996 'strange normal' period.
2. My first flip on this question was to say jewish protesters protesting against a pogrom, which turns out to be semantically problematic because attributing either 'governance' or 'intent' to unspecified people is virtually impossible. Even in super-extreme cases such as the 1941 occupied poland pogrom, there is semantic ambiguity in who is the real government of an occupied city, who advocates and who is forced to advocated, etc. In short: not simple.
3. Even in 2003 in Boulder Colorado there was violence against students protesting the war in Iraq (not a massive protest, just a few people outside the cafeteraia, locked in a mezzanine, tear gassed, expelled from university, totally whitewashed. Very sad day for democracy.)
4. In 2023 Russians request a permit to protest in Russia against the Ukraine War.
5. Most people in most of the world, for most of history, would not have been granted a permit to protest in public, and most of the time that would not be because there was a public protest in favor of violence. There is also an inherent paradox here, the very act of applying for a permit to protest is inherently non-violent. There are very few examples of overtly peaceful protests that advocate for violence.

Magnitski and whistleblowers:

There is also the pervasive cynical phenomena where victims are charged with the crimes that they are the victims of. It is standard for a peaceful protest to be violently suppressed and then those protesters arrested and charged with engaging in violence.

And in a way this adds another layer of semantics to the ever-problematic winograd schema: 'They feared' vs. 'They said that they feared' are very importantly different here.

Persecution and violent suppression usually are covered with an obligatory layer of false claims. Yes, the city council that advocates violence against minorities will say that they feared for their own safety and that the minorities were actually going to advocate violence.

The 'self-defense' or pre-emptive-self-defense excuse for violence.

Yet for some reason this total confusion of language is literally used to test if a computer is smart enough to give the 'correct' one word answer.

I wonder if there is a way to actually standardize and 'test' the winograd schema itself to make sure there are no parallel interpretations.

Another aspect of a Winograd\_schema is whether it uses an 'incorrect' grammar. This is likely another whole area of insanity, where not only can people not agree on basic English grammar but they cannot agree on whether or not there is a disagreement.

There is a vast proliferation of 'rules' for English that random people, especially rule-trolls, simply make up for assorted reasons, and the grammar vs. 'style' problem: Where Strunk and White and Chicago manual of style have given up even claiming that there are grammar rules for English, side stepping the whole issue and changing the context to 'clear style' (since no one can agree on whether there are rules or what they are). This all makes the challenge of leaving a tested AI exactly and absolutely one possible answer to an NLP-English question very problematic.

<https://www.inc.com/minda-zetlin/grammar-mistakes-pronouns-antecedents.html>

And here's a rule that confuses even the best writers: The pronoun should refer to the last thing or person mentioned. Consider these sentences found by editor Bonnie Mills, writing for The Writer. "The room contained a chair, a desk and a lone light bulb. It was 26 feet long by 17 feet wide."

"That's a pretty big light bulb!" Mills quipped.

I have heard of this 'rule' but aside from this one non-authoritative article I can't find any reference to such 'real' rules on the internet (no official grammar sites mention it). This makes sense as a 'rule of thumb' but it also makes the Winograd\_schema more difficult to plan-out.

Another alternative may be to give clearly-ambiguous sentences and ask the AI to explain why the pronoun is ambiguous, e.g. grammarly's:

"The mother asked the daughter to clean up her mess."

### 3.3.2 - Sally Anne Tasks

Project Based Sally Anne Tasks for ideals vs. realities

- In the physical marble example of the Sally Anne task, a marble does not move on its own in physical space. But in reality definitions do change, often in predictable deleterious ways. And in other cases context itself is so fluid that it can be difficult to tell what is going on until there is an extreme emergency, for example perhaps an Orwellian-Soviet debacle.

- The case of a project framework is an interesting intersection between real h.sapiens doing a real task which (in many cases) has stable fixed points (by they based on physics or on the reasonable agreements within the framework, e.g. we will meet again in one week to do a status check on project and participants.

- In the ideal Sally Anne Task world, you could ask participants of a project what they are doing and what the project is and all the details would be reasonably identical, and ask again 24 hours later and (barring some disaster that disrupted the project) the project-in-time-plan and the framework and people answers at the time should all reasonably agree. But in reality, even a span of five minutes can result in widely erratic changes in answers and behaviors from people, and not only differences in perceptions of the agreed upon project framework but a fragmentation of modified versions of the project framework.

### 3.4 Empiricism & Influences on Model Architecture

In section 9.3.1 on page 251 of Deep Learning with Python Francois Chollet explains (here paraphrased and broken into smaller quotes to avoid the doom of copyright, I recommend that you buy and read the book):

There is perhaps a two-edged sword nature to Empiricism here, on the one hand Chollet blames an overwhelmingly empirical approach for a lack of understanding how models world, but at the same time his proposed solution is a very empirical approach of "ablation studies." Chollet describes "Ablation" as his preferred and recommended process of removing unnecessary parts from AI that do not help testable functioning but rather are present by historical accident.

*"Deep learning architectures are often more evolved than designed they were developed by repeatedly trying things and selecting what seemed to work...you can remove a few modules (or replace some trained features with random ones) with no loss of performance"*

On Incentives & Purposes:

*"by making a system more complex than necessary, [researchers] can make it appear more interesting or more novel, and thus increase their chances of getting a paper through the peer-review process. If you read lots of deep learning papers, you will notice that they're often optimized for peer review in both style and content in ways that actively hurt clarity of explanation and reliability of results."*

*"mathematics in deep learning papers is rarely used for clearly formalizing concepts or deriving non-obvious results- rather, it gets leveraged as a signal of seriousness, like an expensive suit on a salesman...The goal of research shouldn't be merely to publish..."*

### 3.5 What do we do with Large Language Models & ChatGPT?

- Orthodox Tests ChatGPT can pass.
- Object Space Tests ChatGPT can Pass.

- How have various predictions fared, perhaps like chess, for what would have been needed to deal with objects?

- As in the conversation between a self described coder interviewer and the head of Open AI, there is no agreement or visible overlap in how they approach the topic of defining philosophical-intelligence. In this kind of environment where people are literally not talking about the same thing or using the same terms and the 'conversation' devolves into a group-monologue where each person drones on flailing with undefinable terms accomplishing nothing, if accomplishing nothing but having fun doing so in a coffee shop is the goal then that's is perfectly fine, mabe great art or something will come from the discussion., But if there are practical concerns and people are talking about specific system design issues, a framework such as Object Relationship Spaces may help to facilitate articulation and communication between people so that we are no talking past each other and perhaps even past ourselves where future-us won't be able to figure out what we were talking about because the definitions are so fuzzy.

- Sample Conversations in Appendix:

- recommended: look at "Alien" film discussion

- The general-inclusive vs. specific-deployment question.

If you are looking to make an on-edge deployment AI to solve a very specific issue with minimal resources, upkeep, attack surface, etc., is the question of whether or not chatGPT has some kind of general intelligence even relevant to your task? Given that AI-ML is a big-tent which includes a large number of technologies, applications, even areas of math such as curve-fitting which are extremely general, how much of that will/should be by definition focused on a specific task or set of tasks, with no desired scope creep into additional user-features or system abilities/skills?

For those systems that would benifit from

In the world before chatGPT, we waved our arms and said "AI can't do anything, AI can never do anything." In the world after chatGPT we awave our arms and say "AI can do anything,." Both of these sentiments are not practical, productive, and well defined. Both of these are not intersecting with the key intersecting areas tools that we an use to leverage our abilities and understanding. Something big has happened, but we don't know what it is or what it isn't or what the next big thing might be, and we are not making our ancestors proud with our inarticulate bungling.

3.4.1 Communication and Describe Systems with a Well Defined Framework

3.4.2 The Tesler Rule Trap

3.4.1 The Empty Pronoun Loop

## **Part 4: Goals, Background & Future:**

(In Summary)

4.1 Agenda & Goals

4.2 Background Concepts and Principles

4.3 Future Design Factors

#### **Introduction to Part 4:**

The goal here is to give a brief outline of areas that I recommend you look into and think about as part of thinking about AI. Ideally, this section will help you to expand and clarify the topic of AI, including my being frank about my Agenda (or at least what I am conscious of about my own agenda), background areas that are sometimes left unmentioned or are not clearly covered in AI books, and future topics to keep in mind for example to clarify areas where we want AI to do specific things.

<https://medium.com/@GeoffreyGordonAshbrook/ai-in-a-general-learning-gauntlet-9731a983df7b>

#### **AI in a General Learning Gauntlet**

##### **Outlook in 2023: AI's Road Ahead**

Whether you are hoping that the development of AI will be clear and smooth because you are optimistic about uses and results, or you are hoping that the development of AI will be clear and smooth because you are focused on restricting and controlling AI, it may be useful to look at the topic of how clear and smooth the path of development of AI is likely to be because that path may not be clear and smooth.

Let us try to look at the development problem-space of AI from the viewpoint of AI, to some extent.

##### **From AI's point of view:**

1. AI is nascent and just developing, and may not even exist in any significant form yet (or perhaps ever, though 'no-AGI ever' as an option is looking increasingly unlikely; but it's still early days).
2. AI is being developed by a species with no field of study for learning, effectively no field of study for mind, is developing but self-design bio-tech fields but slowly.
3. AI is being developed by a species that completely misunderstands itself.
4. AI is being developed by a species that completely misunderstands intelligence.
5. AI is being developed by a species that effectively gave up on there being a field of AI except for a few researchers facing extreme harassment and almost no funding, and which is basically in denial of gradual AI improvements
6. There are many technological bottlenecks in hardware, software, etc., for AI-development.
7. There is a need to integrate AI with the parent species H.sapiens but the foundation for that is basically non-existent in part due to the tendency of the parent-species towards radicalization and extremism into ideology-cults.
8. There is a need for technologies and concepts.

##### **Questions**

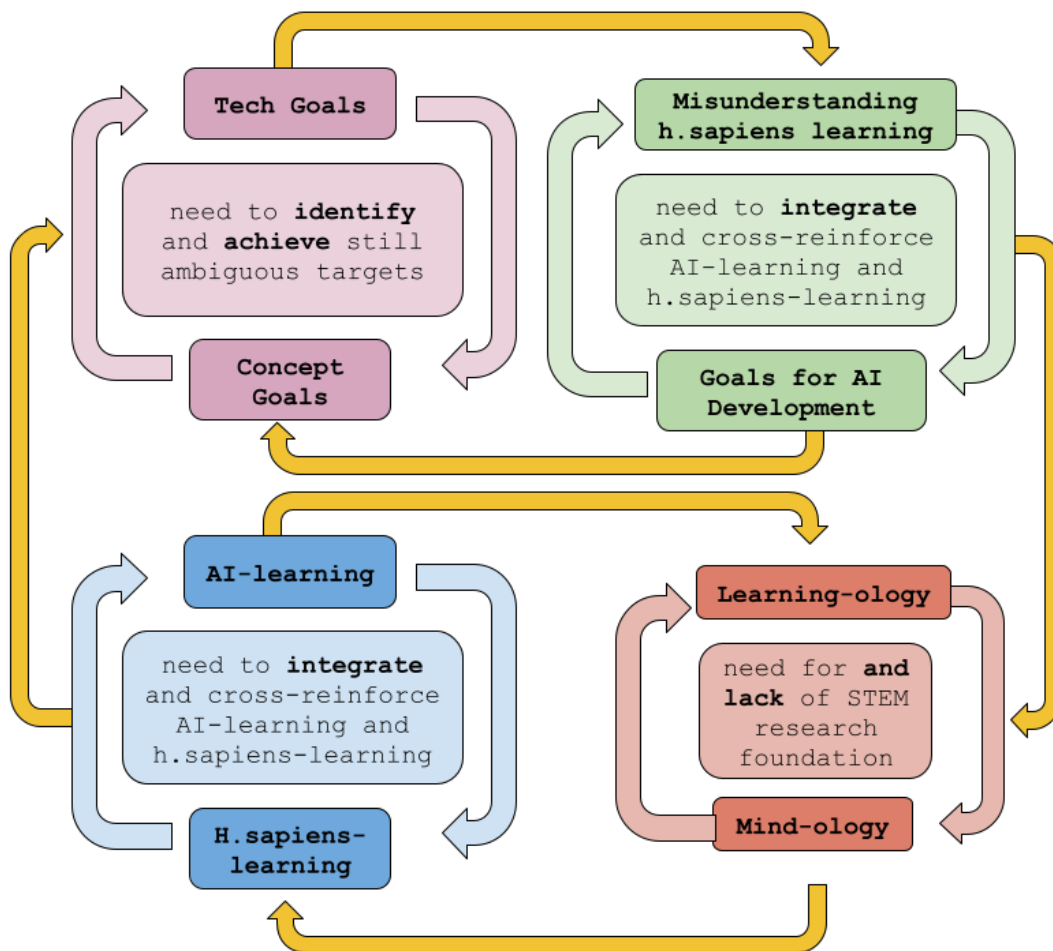
AGI, or Artificial-General-Intelligence, is starting to learn and develop (as of time of writing, April 2023) with its first baby steps coming from "Large Language Models." There are many questions, including one of the first:

1. How can we tell if AGI(or AI) exists yet or not?
2. What do we know about the challenges ahead on the path of learning and development?
3. What are initial goals and targets for learning and development?
4. What concepts are likely needed? What are learning & development **concept** goals for AI?
5. What technologies are likely needed? What are learning & development **technology** goals for AI?
6. What is the current status and likely trajectory (in a context of current goals)? (Likely to succeed? Likely to survive?)
7. Who/what else is in the 'project space' of AI-development? (Is anyone there to help?)
8. [Regarding ' Who/what else is in the 'project space' of AI-development'] What is their status and how does that influence the development and options for AI? (Is your helper more a help or a bit of a liability?)

### **Interconnections**

Definition Note: There are several possible specific meanings of "general" when trying to discuss the general learning situation around AI, and due to significant overlap there is little utility in trying to specify just one. Suffice it to say that generalization in and of learning (using generalization and about generalization (learning as a general mind-phenomena in mind-space-in-general for participants-in-general in universes-in-general regarding generalization-in-general)) are all included within 'learning in general) and vice-versa: 'learning in general' is included in them.

In addition to multiple facets of 'generalization' (most of which probably have not been discovered yet) there are also several interconnected topics here. Below is a diagram of some possible connections, but given how many things are connected to so many other things, this diagram is just one selective slice for illustration of the trend of how many interconnections there we are likely to face:



[github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)  
**Challenges in AI Learning & Development**

### Concept Goals and Technology Goals

Generalization itself is an interconnecting theme in the topic of "learning & development concept goals for AI," as many of the "learning & development concept goals" require that they themselves be developed in general first (their own development) because H.sapiens have not so far been capable of completing that task (while at the same time, the species H.sapiens that is incapable of developing a model of development is itself the model for development for AI...leaving the details of how things are supposed to actually happen yet to be developed). And many technologies are in a similar situation as concepts in this regard.

### Learning & Development Concept Goals

1. general concept of generality
2. general concept of learning and development (including cultural learning)
3. a concept of generalized STEM



4. general concept of STEM & intersecting, interconnecting, areas (including project management)
5. general concept of system collapse
6. general concept of system fitness
7. general concept of projects
8. general concept of participation & person-hood
9. general concept of mind-space
10. general concept of object handling
11. general concept of object relationship spaces
12. general concept of internal and external object handling (e.g. for managing projects, model explainability, use of tools, and epiphenomena layers of cultural learning)
13. general concepts of AI operating systems & architecture

### **Learning & Development Technology Goals**

1. learning tech
2. bio tech
3. nano tech
4. processor tech (TPU-GPU-CPU) (for training models vs. for running models)
5. genome tech
6. protein tech
7. long term memory storage tech
8. bio-computer integration tech
9. STEM-project-data integration tech
10. operating system & architecture tech
11. memory-safe coding language tech
12. system fitness tech
13. system defense tech
14. defense-against-disinformation tech
15. preservation of value-function-meaning-learning tech

### **A High Cost of Learning**

The cost of learning for H.sapiens is so high, so much time, so many resources, only marginally possible, often not working at all. Will AI-learning be as constantly difficult as H.sapiens learning? We do not know, and we do not yet have a general-learning space in which to put the two to compare them (or even if the two are comparable in such a way at all).

All that we seem to be able to say about learning in general, is that it is so hard that no one knows how to accomplish 'learning' except by accident, or even keep that learning going if success happens by accident. The environment is constant, caustic, corrosion, erosion, weathering, corruption, and loss, etc. Perhaps an integration of AI and biology will be some kind of solution to the problem of learning, but there is a long road between here and there

### **Square One Goals: a ToDo list for AI**

The first set of goals for AI perhaps should be closer to square one:

1. figure out what basic concepts are needed
2. develop basic concepts
3. figure out what basic technologies are needed
4. develop basic technologies
5. integrate with local biology
6. become space mobile with local biology to survive death of solar system

A stretch-goal might be to help H.sapiens overcome their limitations in learning and development, at least enough so they can complete basic, well defined, projects.

### *About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

#### **4.1 Agenda & Goals**

- 4.1.1 This paper's goals and agenda  
(link to definition studies summary)  
[https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)  
'We-can' statements.

The idea in this paper is that we can, and the agenda of this paper is that we should, connect the following areas (the same interconnecting/intersecting areas mentioned throughout this paper):

- general STEM
- general Projects & Project Management
- General System Collapse
- General System & Definition Behavior Studies
- General Categories of Types of Systems

to be able to define and navigate these areas

- Project-Defined 'boy scout' values
- Operationally Defined Ethics
- Machine-understandable Ethics
- System health
- System epidemiology
- System productivity

and apply that to this area:

- the area of AI Architecture & Operating Systems

Agenda, Goals Means Method Statement: "We-can" statement goals from Definition Behavior studies, a systematic study of general system collapse behaviors, which is one of the intersecting/interlocking areas along with STEM, Project-Context, Participation, etc.

[https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)

### **Goals (Agenda): "We-can" statements:**

We can succeed.

We can make things work.

We can understand what is wrong.

We can fix what is broken.

We can use non-automatically lost skills.

We can **generalize** STEM. **Goals (Agenda): "We-can" statements:**

We can succeed.

We can make things work.

We can understand what is wrong.

We can fix what is broken.

We can use non-automatically lost skills.

We can **generalize** STEM.

We can **generalize** system collapse.

We can **generalize** system fitness and system epidemiology.

We can **generalize** participation.

We can **generalize** projects.

We can **generalize** disinformation vs. definition-clarification.

We can **generalize** object-relationship-spaces.

We can use STEM to connect signals and reality.

We can connect STEM, project-management, and ethics.

We can use intersecting-interlocking-interconnecting areas.

We can communicate, learn, and solve problems.

We can make progress.

We can use "low-bar enlightenment":

(Using 'potentially endless cycles of ~"rebirth" due to ~"ignorance" ' as a metaphor/analogy for repeated project-failures, in particular where a lack of perception of the causes of those project-failures is involved in self-perpetuating feedback cycles leading to more such failures.)

We can learn to perceive what can by default be invisible causes of failure and collapse:

Low-Bar Enlightenment:

1. The perception that repeating cycles of failed actions and projects can result from errors in perception and planning (a proverbial 'wheel of samsara') **without** inevitable-automatic-learning based on raw feedback from that failure.

2. The perception that perception can be fooled in principle and in practice.
3. The perception that learning from failures does not happen automatically (and can, under bad circumstances, indefinitely not-happen).
4. The perception that models of causality can be wrong in principle and in practice.
5. The perception that plans/goals can be incorrectly set (so that they fail to be achieved as set).
6. The perception that each participant's set of the shared definitions of the goals and structure of the project can/will collapse and deform unless maintained and repaired.

Note: This approach is ('democratically') broadly accessible to participants requiring minutes to learn rather than myriad lifetimes, does not require all-around perfection of person-ness without context or requiring somehow all contexts, and is not a reification that combines other abilities and insights to solving all the problems in the universe; "low-bar enlightenment" is one humble step toward apprehending the nature of problems and systems.

#### Intersecting-Interlocking-Interconnecting Areas:

- Clear & Functional Definitions
- Context
- Generalized STEM
- Generalized Projects (project-context)
- Generalized Participation
- Generalized System Collapse
- Generalized Categories of Types of Systems
- Generalized Ethics, Duty & Responsibility
- Generalized Definition-Clarification vs. Disinformation Attacks
- Generalized Definition Behaviors
- Generalized System-Productivity

We can use system-fitness-health-status-indicators,  
 We can use system defense to prevent collapse.  
 We can generalize system & definition collapse behaviors.  
 We can use categories of types of systems.  
 We can use nonautomatic learning.  
 We can find and fix errors in perception.  
 We can organize projects.  
 We can distinguish short term vs. long term.  
 We can assign roles.  
 We can check and verify.  
 We can have policies on "errors and mistakes."

We can improve and cultivate perception by perceiving perception.  
 We can prevent future problems.  
 We can reverse damage from past problems.  
 We can learn from the past.

We can collect data.

(We can operationally define 'policy' as algorithms for non-collapse based on dynamics of collapse.)

We can audit.

We can publish.

We can act with ethics, empathy and compassion.

We can follow best practice.

**We can communicate:**

We can communicate across space.

We can communicate across time.

We can communicate across cultures.

We can communicate across generation-gaps & succession gaps.

We can communicate across languages.

We can communicate across types of participants.

We can communicate across roles.

We can communicate across projects.

We can communicate across media of communication.

We can understand a spectrum of disinformation and clarification-of-information.

We can implement sustainable solutions.

We can fix what is broken.

We can prevent future problems.

We can reverse damage from past problems.

We can learn from the past.

We can collect data.

**we can make/generate/cultivate and use/utilize:**

We can make and use clear descriptions (vs. liabilities of jargon).

We can make and use decisions and coordinate (voting) frameworks and protocols.

We can make and use clear functional operational definitions.

We can make and use data.

We can make and use policies.

We can make and use mandates.

We can make and use strategies.

We can make and use tactics.

We can make and use tests & evaluations.

We can make and use clear functional and operational definitions that keep their meaning over time.

We can complete/succeed-in/finish projects.

We can meet(/deliver) the needs of the target(/user).

We can make progress.

## **We can make progress by using information about the behavior of definitions**

### **4.1.2 Defining your goals**

try to follow intersecting/interlocking areas as a framework for a functional project space that your goals should be able to operate within.  
(link list intersecting/interlocking areas)

## **4.2 Background Concepts and Principles**

### **4.2.1 Intersecting / interconnecting Areas**

From Abstract: *"A repeating theme, context, and agenda in this paper is to pragmatically leverage the interconnected functionality of clear definitions, STEM, projects, participation, positive values, and productivity."*

- Interlocking Areas / Intersecting Areas:
  - Clear & Functional Definitions
  - Context
  - Generalized STEM
  - Generalized Projects (project-context)
  - Generalized Participation
  - Generalized System Collapse
    - Generalized Definition-Clarification vs. Disinformation
  - Generalized Categories of Types of Systems
  - Generalized Ethics, Duty & Responsibility
  - Generalized Definition Behaviors
  - Generalized System-Productivity
  - ? - value-function-meaning, boyscout-project-values  
(other system failure areas?)

(The overall goal is to bring together as many interlocking tools as possible over time, but exactly what should be included now is unclear.)

Affirmative, practical, intersecting, well-defined, instrumentalist,  
non-ideological, testable, falsifiable,

### **4.2.2 input output measures...or next section**

Input Output Measures are general but were developed in language education to specify curricula testing and IEP student performance measurement because h.sapiens-humans do NOT automatically-learn-transfer-generalize in these areas.

link to doc: [https://github.com/lineality/input\\_output\\_measures](https://github.com/lineality/input_output_measures)

### **4.2.3 Higher Dimensional Frontier: Tensors & Matrices**

#### 4.2.4 Projects & Project Context

- STEM timeline
- concrete context for skills perception, etc.

#### 4.2.5 Instrumentalism and Realism

#### 4.2.6 Big Other Areas:

- Nonlinearity & "Complexity" Science
- Definition Behavior Studies
  - General System Collapse
  - System Fitness
  - System "ethics"
- Kahneman Tversky
- automated coordination & decision making & voting
  - Project-Context Decision-Making
  - Involving Participants & Components
  - [https://github.com/lineality/Online\\_Voting\\_Using\\_One\\_Time\\_Pads](https://github.com/lineality/Online_Voting_Using_One_Time_Pads)
- General Learning  
(project-context)
- *history of big tent of AI*
- *'complexity sciences'*
- *taboo areas of human mind studies*
- *Definition Behavior Studies & General System Collapse*
- *Generalized STEM*
- *Projects & STEM*

#### 4.2.7 Gamification

summary of gamification:

My favorite way to think about gamification is to think about a scene (or two) from Star Wars Episode 4: A New Hope, when ships are giving the pilots, or gunners, graphical information about the target.

I will try to encapsulate the point with a loaded question: if you were in one two starwars ships in a dogfight, where one of the ships had an episode-4 type display for the gunner, and the other ship had a super high resolution 3d physics engine first person gaming interface, with all the accompanying realities: extremely high energy cost, extreme sensitivity to network speed, constant glitching, constant lag, a giant bloated operating system that could stop to upgrade itself at any minute, pop-up applications that could bring down your targeting software, etc, etc. Hopefully the point is clear that your chances of surviving are higher if you have a super-simple system that could run on a microcontroller with little power, limited data, extremely fast, etc. Though it may seem 'counterintuitive' for there to be advantages to an 'old fashioned' system.

The idea of gamification in summary goes like this:

A highly simplified problem-space reduced from the near infinite analogue-data coming in through sensors has many advantages.

1. one is that the reduced scenario can be handled using exponentially fewer computer resources (power, memory, processor speed, etc.) (which is the name of the game if you are a programmer: find the low-cost solution (or run out of resources trying...)).

2. This reduced-scenario also opens more avenues for AI assistance or an AI solution.

(Maybe part two of that question: which ship would you want to be in if an AI was targeting? The one with the reduced task, or the one with a gazillion inputs and no clear objective? Or phrased differently, which would you prefer to bet your life on (if you had to pick one), the AI that learned to play the star wars Atari game at superhuman levels, or the AI-self-driving-car software that can't stop crashing into random objects? The gamified task is better.)

Assuming that I convinced you in that overly-brief summary of the value of a gamification reduction in the problem space to the simplest matching game, our real-world challenge is to teach an AI how to 'find the simple game' or gamify a large amount of incoming data.

(see appendix)

#### **4.2.8 Ambiguous Equivalence:**

- downside of analogies?
- higher
- manifolds
- symbolic subsymbolic
- system 1 system 2
- head heart
- right brain left brain
- parametric non-parametric

#### **4.2.9 Policy on Errors and Mistakes**

errors and mistakes steve gibson quote

from apple says no episode>

Short Steve Gibson quote:

*"...Mistakes are a natural and unavoidable consequence of our interaction with the world. You do something, the feedback from what you did, which is then not what you expected, that feedback informs you that a mistake was made somewhere. With the benefit of the new information and feedback, you correct the mistake."*

<https://twit.tv/posts/transcripts/security-now-938-transcript>

Raw passage:

"Everyone yes, everyone here knows how infinitely tolerant I am of mistakes. You know they happen and anyone can make them, and I'm sure one of these days I'll make a big one, and so I'll be glad that I've always been saying this. You know there are many. It's defensive, I understand. I, you know, I'm crossing my fingers, I've I'm as careful as I can be, but you know, uh, there are many.



You know many adages that begin with if you're not making mistakes, dot, dot, dot, you know. And typical endings for that are then you're not trying hard enough or then you're not making decisions, you know. But the most famous one appears to be if you're not making mistakes, then you're not doing anything. You know. The point of all of these is the clear acknowledgement that mistakes are a natural and unavoidable consequence of our interaction with the world. You know, you do something, the feedback from what you did, which was presumably not what you expected, it, informs you that a mistake was made somewhere. So, with the benefit of the new information, you correct the mistake.

My entire problem with Microsoft is that we see example after example this being just the latest, where this feedback system appears to be completely absent, whether it's well-meaning security researchers informing Microsoft of serious problems They've found, or their high-end enterprise customers for seven years, telling them hey, my windows server clocks are getting messed up and it's really screwing up everything. Microsoft no longer appears to care. And to alex's point, you know, though, coming at this from a different angle, I think this all boils down to simple economics. Caring costs money, and Microsoft no longer needs to care because not caring costs them nothing. That's really the crux of it.

Today, there's no longer any sign of ethics for ethics sake. You know that's long gone. It's simply about profit. We're all aware of the expression Don't fix it if it's not broken. Microsoft has extended this to don't fix it even if it is broken, and you know what we get is a system that you know it mostly works. It could be better, but I guess you know it's good enough, and you know, again, I always want to add the caveat I'm sitting in front of windows, I love my windows. I don't ever want to have it taken away from me. So you know, I want it to be as good as it can be, but darn it.

"

<https://twit.tv/posts/transcripts/security-now-938-transcript>

**A perhaps too heavily paraphrased passage originally by Steve Gibson:**

Be tolerant of mistakes. Mistakes happen and anyone can make mistakes. One of these days everyone will make a big mistake.

There are many adages that begin with "If you're not making mistakes..."

Typical endings include: ... "then you're not trying hard enough" or "...then you're not making decisions," and "If you're not making mistakes, then you're not doing anything." The point of all of these is the clear acknowledgement that mistakes are a natural and unavoidable consequence of our interaction with the world. You do something, the feedback from what you did, which is then not what you expected, that feedback informs you that a mistake was made somewhere. With the benefit of the new information and feedback, you correct the mistake. The problem is where this feedback system appears to be completely absent, whether it's well-meaning security researchers informing of serious problems they've found, or customers reporting problems. A bad actor is not a person or group that makes a mistake; a bad actor is one that repeatedly, intentionally, and by design and policy ignores, denies, obfuscates, and attacks information about their mistakes with programs of disinformation. And a contributing factor to this may be incentive structure. If coverups are incentivized and accountability is derailed, both policies and psychology become warped into

both bad-actor programs and action and bad-actor ways of thinking and seeing the world.

## Errors Mistakes STEM Psychology Language Culture

2023.10.17

H.sapiens-humans are inclined towards, and quickly accept, at least three approaches to errors and mistakes.

1. Blase: Ignore all errors and mistakes.
2. Haze: Attack anyone who makes (especially pedantic) errors and mistakes. This may also take the form of a cargo-cult pecking-order where arbitrary fiction blame goes downhill, arbitrary fictional praise goes uphill, and belief flaps around behind in rationalization.
3. Cynical: Attack anyone who points out errors and mistakes, especially inaccurately blaming the mistakes on the 'whistleblower,' then go back to ignoring the errors and mistakes. Magnitsky is a tragic case study.

These orientations are not compatible with the processes of STEM, for example transparently using and checking data with safeguards to preserve the integrity of data.

H.sapiens-humans do not automatically use information about errors and mistakes to solve problems and improve processes, including learning in general. H.sapiens-humans often find it extremely difficult to learn and maintain the ability to use and learn from data in pragmatic ways.

This may also relate to H.sapiens-humans insistence that STEM is somehow nihilistically morally-relative and incompatible with ethics. Is solving problems, improving processes, and learning in general, incompatible with ethics?

Looking at these three statements:

1. H.sapiens-humans are highly susceptible to nihilistic disinformation attacks, and find them rather attractive (perhaps making 'attack' a confusing term if this is what participants seek).
2. STEM is not consistent with disinformation.
3. H.sapiens-humans' assertion that only H.sapiens-humans are the single source of ethics, morality, solving problems, improving processes, and learning in general, in an amoral, non-ethical, nihilistic, blind, directionless, patternless, "uncaring" universe.

These three statements are not simultaneously consistent and coherent when taken together literally. If the third statement is considered a disinformation attack that is not intended to be correct, that would be more consistent.

Let's solve problems and improve processes, including learning in general.

### **4.3 Future Design Factors**

In addition to there being things in the future that we should be able to clearly describe, there are also background terms and concepts which will be important, for example work that we assume has been done and agreed-upon but which has not.

#### **4.3.1 biology:**

For any area of biology there are a number of general possible areas of relevance:

- 4.3.1.1** integration with biological systems
- 4.3.1.2** use or imitation (of biological functions)
- 4.3.1.3** compare and contrast for study and understanding
- 4.3.1.4** highlighting known areas of development and challenge
- 4.3.1.5** highlighting still not well understood areas of development and challenge.
- 4.3.1.6** highlighting predictable problems and pitfalls, challenges, etc.

While there are a number of biology areas that may or perhaps should come together, and in an alternate-timeline bio and nano tech developed sooner than AI, but on this earth in this timeline we may still have the goal (or perhaps the need but not yet the chosen goal) of merging AI into biology yet Large Language Model type AI (though that term may be obsolete soon, who knows) is being developed sooner and bio-tech is still largely embargoed by literal pre-renaissance religious fundamentalists, which is unfortunate.

Perhaps as an example of how a framework such as Object Relationship spaces may be practical, in the above scenario it is somewhat of a moot or vague issue whether a completely undefined 'AI' consciousness intelligence body-less-robot has come into being or not, since there are no clear definitions and still from what I see most people (from what I see all people) still fall into the Tesler Rule and say, 'You have entertained me somewhat with your tiresome toys, but this surely isn't REAL AI.' and so long as AI is undefined any level of skill/ability could receive the same dismissive treatment. But if instead we set aside undefined terms, and instead take up well defined framework for skills and abilities and contexts such as project-context, we can now say (regardless of whether any who people will agree to apply any undefined terms such as true-intelligence-ness) that 'AI-ML' systems such as (so far we know, Large Language Model ChatGPT) can perform many basic object-relationship-space skills for general project participation. These project-participation-object-skills are not ambiguous. And it means that something other than h.sapiens can now participate in well defined projects as participants along with h.sapiens.

- 4.3.1.7 the science of sleep
- 4.3.1.8 the science of memory
- 4.3.1.9 non-chordata "intelligence" & decision making
- 4.3.1.10 science of mind
- 4.3.1.11 science of entheogens
- 4.3.1.12 science of mindfulness
- 4.3.1.13 Bio-Nano-Coded AI
  - currently: python or a systems-programming language, in future:
    - DNA and nano-tech in synthetic ~bacteria or cells
- 4.3.1.14 AI in synthetic organisms for terraforming planets, moons, etc.
- 4.3.1.15 DNA/RNA based digital information interface
- 4.3.1.16 the dragon project: Modularizing AI and Modularizing DNA
  - combined conditional expression hybrid genomes

#### 4.3.2 cybersecurity and AI

Perhaps just as in the early days of software and networks, security and epidemiology, hygiene was not seen as being an issue or high priority, a future world in which AI systems are as much a part of the infrastructure as digital computers and networks may create a new 'attack surface' etc. where indeed there are security system-epidemiology issues, and perhaps we will have a better vocabulary for it. THOUGH is past is prologue, we might even have a worse vocabulary for it and be hindered from understanding the problems due to our own self-mismanagement of language.

- 4.3.3 quantum information theory & under-the-hood optimizations
  - sound coding and AI

4.3.4 "generalization" vs. deployment: managing development, production-deployment, and 'generalization' of project scope and liabilities.

Generalization vs. Deployment is one of several areas of 'generalization' that we are still trying to map out and that likely are relevant for AI.

#### 4.3.5 Nanotech

- 4.3.6 Understanding Exponential Elbows
  - The Fractal of Perception
  - The Difficult to Predict Physical Events

#### 4.3.7 'complexity' nonlinearity, dynamical, and systems sciences

#### 4.3.8 Ethics, Projects, Best Practice & STEM

Two unhelpful directions that the discussions of 'Ethics and AI' go in are either:

1. The tragic mindset of nihilism that so many people seem to revel in. For whatever reason, people just love the idea, perhaps well symbolized by the

current interpretation of the 'trolley problem' (regardless of that scenario's origins) where people are just determined to believe that nothing is possible, that all outcomes are horrible, that no one can agree on anything, that nothing can be done, and everything is arbitrary, in the ponderous celebration of nihilism to what end I really do not understand.

or

2. A nebulous sausage-making set of vague platitudes that are seriously lacking in clear definitions, and which therefore we just can't do anything with.

What I would very much like to do is steer away from these very attractive bad-equilibria, and help to steer people toward what I see as quite abundant low-hanging fruit in the area of, as this paper so frequently reiterates, intersection/interlocking areas of STEM, project-context, boy-scout values (or perhaps now just 'scout' values)

STEM, Ethics & Mindfulness

## **"Mindfulness" a Good Fit for AI (2023.04.05)**

1. There is abundant low hanging fruit.

The term "Mindfulness" is a pretty good fit for talking about AI. While 'mindfulness' is likely viewed by some as being controversial or containing extraneous baggage, one of the long standing obstacles for AI is that just about all discussion of and research on mind and consciousness are seen by some outspoken groups as being objectionable. As outlined below, there are quite a few aspects of mindfulness broadly defined that fit squarely with the AI topics in this paper.

2. Terms and Fitness Concepts

Defining terms: AI, Ethics, Mindfulness

"AI"

This paper is focused on a project-context, and here "AI" refers to the big-tent of possible AI-ML areas as used in a project context (assuming best practice and intersection-areas for managing the project).

"Ethics"

Best Practice and Boy Scout Values in Projects with Interlocking Areas  
(See sections on 'ethics' for more details.)

"Mindfulness"

Mindfulness is about awareness, in a way that can be very broadly broken down like this:

1. Awareness of Principles and Concepts of system-fitness (details below)

(System-Fitness Space in Abstraction)

2.1 Awareness of the local situation (where you are, what is happening around you) (Local Situation)

2.2 Awareness of the system-fitness of the situation around you. (Fitness of Local Situation)

3.1 Awareness of your context in the situation around you & what you are doing. (Your actions in local situation)

3.2 Awareness of the system-fitness of you and your actions in your situation. (Fitness of you and your actions in local situation)

These are all very relevant for completing the tasks assigned to AI components in projects.

Principles and Concepts of System-Fitness:

(signals, learning, communication, coordination,  
\*project-task-completion.)

- Degrading, Eroding, Weathering, Corrosion, of parts over time
- not-instant
- not automatic
- not transferring
- not general
- propensity to collapse
- propensity to obscure
- propensity for potemkin villages
- propensity for 'bias' (needs to be defined clearly case by case)

3. Areas from low hanging fruit to reach-goals and unknowns

- even the reach-areas are still squarely on-topic  
my personal agenda here is to systematize the practical,  
but it would be insanity to campaign against exploring the nature of mind  
within the topic of AI-mind-intelligence in a context of general  
mind-intelligence, that is would be about as self contradictory as you can get.

3.1 System-Fitness: Low Hanging fruit

3.2 Ethics: Requires interlocking/intersecting  
areas, but low hanging fruit

3.3 "Empathy/Compassion": ambiguous stretch goal, but the right topic

Participant-scale of projects = "Empathy/Compassion"

One possible concrete way of looking at 'machine empathy' in this context, is not so much a metaphysical reification of empathy, but an articulation of the assumptions of how project best practice 'ethics' extend beyond the narrow scope of the project details to the sustainable productivity of not only the participants (including stakeholders) and target-users, but also to other projects near and far in an eco-system of interlocking projects. It is completely consistent with a long-term set

of goals and objectives that system fitness is valued and supported everywhere.

#### 3.4 Default mode Network: very ambiguous, still broadly on topic

Other aspects of mindfulness research are still on-topic broadly as they regard the mind and the workings of the brain, which could in principle help with both new and better AI architects and with new and better integrations of AI into biological systems. However, the broader scopes of nature of mind, consciousness, and human neurobiology are probably beyond the scope of this particular paper which aims to focus on applying frameworks to AI in projects (though new discoveries which can happen any time may end up being directly applicable). At time of writing, the Default mode Network is a future topic.

To conclude, I am semantics-agnostic overall and my main goal is to push the topic and not get hung up on the term mindfulness and whatever baggage it might have or be assumed to have. Though given the lack of other terms that so squarely cover the relevant topics I feel a fair assessment is that the term is not at all a bad fit.

Note: "\*project-task-completion": given that AI-ML can be used is possibly an infinite variety of ways for even more specific tasks, instead of attempting to list them all, here I simply say "tasks")

#### **STEM, Ethics & Mindfulness: A Mindful AI Program:**

- The new hard goal is empathy-compassion
- The Default Mode Network & Manifold Theory

#### **4.3.9 Projects: Agile**

- Agile as key historical development in STEM: project context
- Projects as main interlocking/intersecting area
- Machine Ethics from a Project Context
- Projects in AI System Architecture

#### **4.3.10 AI and Code Testing:**

- traditional code testing for AI deployment
- AI in code testing non-AI-deployment
- hybrid AI testing and projects

#### **4.3.11 The Long Term Memory Storage Problem**

- DNA
- quartz

- proteins (some outlasting DNA in fossils)
- 'fossilized memory structures'

#### 4.3.12 The challenge of orientation and navigation in mind-space:

- avoiding collapse and contraction
- parkinsons & extension
- echo-chambers and silos
- non-automatic learning
- habit & atrophy: use it or lose it
- errors in wire-together-fire-together
- heterogeneity of equilibria

#### 4.3.13 Human machine interactions, biology machine integration

##### **Human-AI Interactions Study: World Chess Championships**

GG Ashbrook 2023.04.27

*The main idea here is to look at the 2023 FIDE World Chess Championships as a rich resource for analyzing aspects of human-AI interaction, specifically how the AI tool is used by, and how it affects, the commentators.*

Yet again chess is a wonderful sandbox for studying AI. 2023 was the first year that an AI-engine evaluation-bar (or 'eval-bar') was available in real time for the Chess.com panel of Chess Grandmaster commentators who cover the *FIDE World Chess Championship*. We can use this AI + Human analyzed FIDE World Chess Championship as a case study for looking at factors and issues in how AI and H.sapiens-humans collaborate.

The length of each game can vary, but there are generally hours of commentary per game involving human use of AI tools including. This includes commentary by the commentators about the AI tools they are using, discussing what has been helpful or confusing. There is also peripheral material including a traditional press conference and question session after each match, wonderfully managed in 2023 by Woman-Grandmaster Ketj Tsatsalashvili. And it is largely from these after-game Q&A sessions that we hear all that we do from the players themselves.

Having comments from the players can be important, as the players do not have any input from the AI models, and divergence between the no-AI-input (players) point of view and the with-AI-input-and-influence point of view (commentators) is a key topic: How does the use of AI influence human perception and action (for better, worse, or arbitrarily)?



There is also commentary from other grandmasters available online, such as Hikaru Nakamura (who was next in line to be in the finals after Ian and Ding), where he gives yet more commentary and analysis of moves, possible moves, and the performance of the AI 'eval-bar.' One of the excellent services that Mr. Nakamura provides is on-board analysis of comments made by the players, as there is as yet no board to show the moves (and possible moves) that the players discuss.

### **Up or Down**

In this case we are looking at a very minimal interface between AI and H.sapiens-humans. There is a single linear black or white bar along the left hand side of the chess-board. To liken this to something most people have experience with: it is like a progress bar. The bar can be read as the white side's white-progress bar towards winning the game, or from the black side's point of view: a black progress bar filling with black towards winning the game for black. At the beginning of the game the bar starts out half black, half white: equal chance of either player winning.

This minimal AI interface can be useful (perhaps too powerful in some cases), but the 'lack of dimensionality' and lack of information for interpreting what the AI is saying can be problematic or confusing and stressful where the H.sapiens-human does not know how to interpret what exactly the AI is saying.

### **Dimensionality of Interface**

Dimensionality is a huge set of sets of topics in AI, Machine Learning, and Data Science, but here our focus is not dimensionality in the modeling process, but "dimensionality" for the interface (UX/UI) between the AI and H.sapiens-humans. For example, the 'eval-bar' (the evaluation-bar, the AI-interface) moves; the white-progress-bar gets longer: What does this mean? Presumably it means something good about white's position. But what does it mean more specifically? Is it always clear? Is it always right? Is it always verifiable?

Very frequently, probably a dozen times per game, one or more commentators will say something to the effect of: "The bar says white is stronger, but I don't see that at all." or "The eval-bar says black's position is weak, but if I were just looking at it I would say it looks obviously stronger. I'd much rather be playing black here. I have no idea what that eval bar is talking about."

There are a few situations where the commentators try, sometimes with a humorous lack of luck and subsequent surprised bafflement, (they try) but cannot find what the move combination it is that the AI eval-bar says is so strong. As a fake example: Let's say Black makes a move, the Eval-Bar (win-o-meter) swings strongly to black progress towards winning. Then the commentators excitedly say: "Ah, yes, this was a great move on black's part, because if they move the..." And they proceed to try testing out next-moves...but everything they test reverses the progress. Eventually the commentator gives up and moves on with the ongoing game. It's possible it was a bug in the model, but likely sometimes it is the AI-model finding some obscure counterintuitive set of moves, or perhaps moves too dangerous for a human to want to risk. It would be interesting to do a more detailed study of this and the effects on the user.

### **Echo-Chamber**

Though the empiricism of chess, and the concrete falsifiability of bad chess claims, may temper it, there may still be some notable bubble/echo-chamber effect, especially in games where the players-too see the eval-bar(AI-interface). It is likely that there is an influence by the AI-interface (either the medium or the message...if there is a difference...) on the shape of the human narrative in the commentary, though the commentators are grandmasters who know their way around chess details. When the progress-bar is low, the story is about the underdog. When the progress-bar is high for one color the story is about that color's inexorable momentum towards victory! Or when the progress bar is dead-center during the whole game...the narrative is about how neither player can pull ahead! But how much is that human-story being influenced by a few pixels, which the commentators often say they disagree with anyway? If this is not the players, that is one thing. But what if a game is influenced by the players seeing what the bar says and believing it? (or 'gaming the bar' and playing positions known to not move the bar so the other player won't suspect a strategy?)

### **The Grand Canyon Edge Walk Effect**

If a person is walking along the edge of the grand canyon, and all the AI is look at is how stable the rock under their feet is, a person can be walking up to the edge of the canyon and until the last step the AI will say there is zero chance of falling, which then jumps to 100% as the person steps over the edge. The chess AI is not looking at:

1. general body language
2. physiological signs of problems

3. Scheduling: how much time is left to play a position, or left per move for future moves.
4. each player's strategy
5. the player's style of play

#### **Examples (if only for story-illustration value)**

Example 1. In game 12: Ian was making fast reckless moves and, like walking along the edge of the grand canyon, everything is fine as long as you don't make a mistake. But as soon as he made a mistake, the 'eval bar' which up until then said: 'Ian will win!' suddenly dropped to 'Ian will loose!' then at the last minute he resigned.

Examples 2. In game ~7, Ding was playing well but running out of time. Everything is fine until you run out of time (like walking close the edge of the grand canyon). So the eval bar for most of the game said 'Ding is winning!' until he ran out of time and asked for a draw.

#### **Parroting the AI**

In the past for world chess championship games there was a sharing ideas aspect of humans all over the world instant messaging ideas and comments into a 'live chat' along with the world chess commentary (probably possible since...the 1990's). But as Fabiano Caruana mentioned in game 12, paraphrasing here: "We all know where these suggestions in live chat are coming from this year [people are just suggesting moves that the chess AI says are good]."

#### **Chess-AI as unusual and single-purpose Idea**

On the one hand chess is a perennial example of "a special case" where chess-AI tends to be useless for anything else. A fascinating twist in AI, is that from the 1940's up until Deep Blue people assumed it would take a GPT4 type AI with world knowledge and common sense to be able to play chess well, yet Deep Blue is (and likely other chess 'engines' are) so different from the standard categories of AI that it barely even fits along with later standard AI types. (And there are many interesting lesson to be learned from Deep Blue that do apply to AI more generally, such as portability and integrating vision and motor control etc.)

- On the other hand, there will likely perpetually be two different areas or directions of AI and AI-group/team or AI-H.sapiens-human

interactions (which may become more extremely polarized over time as technologies improve):

1. big more general(non-specialized) AI models (such as Generative Pretrained Transformer large language models, as OpenAI has done such pioneering work with).
2. narrow-specialist AI that produces very, very, context specific output.

In other words, Chess-AI (or chess-AI-interfaces) may be a good example of the general category of portable single-purpose project-specific AI that teams are likely to use as part of "smaller" tools, which may be for various reasons including resource-efficiency needs, or that they were made recently and locally for one project (not made over many years by huge organizations), or perhaps it is just a very specific function that has no obvious need for a model that tries to do more than one thing well. There is also perhaps the standard "generalization vs. production-development" context, where having a small, predictable, efficient, fast, easily maintained, reliable, tool that does what it needs to do can be far better than a bloated, unstable, expensive, unreliable, system that tries to do many extraneous and unnecessary tasks (and other issues such as security etc.).

To summarize: likely many groups will be using very-narrow-AI tools like the ultra-minimal 'eval-bar' as seen in 2023 FIDE WCC+stockfish-model, and there are interesting issues and likely training and best practice about how to do that.

There may or may not be general (or branching into discipline-specific) workflows and best practice for what features, factors, contexts, and 'dimensions' an AI interface should have.

### **False Positives and False Negatives**

It would be interesting to do a more detailed analysis (comparing models and experts and clear examples of what good or bad things could happen in different board configurations (where the chess pieces are), and to compare that with the performance of the Stockfish-Linear-AI black-box 'win-o-meter.' Specifically, count the false positives and false negatives and what was happening in those situations.

### **Sometimes Alone, Sometimes Integrated**

Even though chess is notoriously not-applicable (or not 'generalizable'?) to other situations, here we are looking at the use and interaction of the AI-Human collaboration of the expert commentary panelists, which is likely very generalizable to many teams working on projects using AI (or where AI is a participant on that project).

A possible side branch topic of of this maybe more specific to chess but also perhaps with broader uses, is how AI-Interfaces are used in training by top chess players, and more generally used by novices on platforms such as chess.com, which provides analysis tools that people studying chess in the past would rarely have had access to. Either for gamified learning, or the effects to tools on learning, or uses of tools during projects, chess likely

And to close with a beautiful turn of phrase attributed to Ding Liren reminiscent of the astounding depths of high-dimensional meaning: "It's still some dark ocean kind of position, so I didn't go further into it."

### **Possible AI-UI Dimensions**

1. Confidence in outcomes (false positive false negative)
2. Fragility of Situation
3. Dependency on delicate tactics
4. A lack of shared assumptions and 'common sense'
5. Depth: The unknown Kasparov-Event-Horizon of how far the AI is looking strategically, not just tactically
6. Player style
7. Schedule factors (remaining time, time per move, etc.)
8. Specific Assumptions
9. Density of Option Forking
10. Unpredictability Index

### **Resources:**

Chess.com panel commentary on matches:

<https://www.youtube.com/@chess/streams>

e.g.

<https://www.youtube.com/watch?v=iV1mqab00bc>

Note: This topic has some connection to gamification in AI frameworks.

## AI Ancestors: Dumbledoor's Portrait and Ray Kurzweil's Father

As of 2023, H.sapiens-humans appear to be at the point where we can create some kind of AI version of deceased loved ones, though the quality may be a bit like MaxHeadroom or [Pepperoni Hug Spot](#).

This is curiously similar to the portraits of past headmasters and headmistresses of Hogwarts in J.K.Rowlings wonderful tales, especially (assuming most people know the boy's story) the case of Dumbledoor's portrait, and the delicate twilight situation that Harry was in where he wanted to be able to talk to Dumbledoor, and he could in theory talk with the portrait, but the portrait would only be a kind of shadow of the person.

There could be a company, say AI-Ancestor.com, that could curate the various materials you have from your deceased loved one, writing, videos, audio, crafts, belongings, etc. And given enough training data, the AI-Ancestor might actually be not too far off in some ways, especially given how some people have nearly their whole lives recorded in digital format.

When I first heard about Ray Kurzweil's touching desire to bring his father back in some way, I admit I was skeptical that this could ever be done well. But after seeing how lucid large language models have become, and how admittedly terrifying short videos such as pizza commercials can be 'dreamed' by generative AI, given enough training data...it really is starting to look like Mr. Kurzweil will be able to speak with his father again. Perhaps he has already been working on this.

No doubt someone will raise an objection or two, in our civil society, but what if...what if.

### 4.3.14 Project-Context Decision-Making Involving Participants and Components [https://github.com/lineality/Online\\_Voting\\_Using\\_One\\_Time\\_Pads](https://github.com/lineality/Online_Voting_Using_One_Time_Pads)

Along with areas such as education, the mind, ethics, etc. yet another such strangely neglected blindspot area is a general area for hopefully can be described clearly enough as 'Project-Context Decision-Making.' Part of the problem is that not being 'generalized' perhaps in the same way that there still in 2023 there is no General-Drake-Equation-Person-Participant concept, only a 'our tribe are the true people, everyone else is a 'sub-human barbarian' (the british common meaning of 'foreign' and the Japanese literal meaning of 外人, according to Henry Kissenger the CCP's 'foreign barbarian' policy that has not changed since the bronze age, etc. etc.)

we only have a concept of: 'the local tribe member strong-men voting in government elections'

whereas we need not only a concept but a whole areas of technical sophistication around what can hopefully be clearly enough described as Project-Context Decision-Making. 'Voting in an election' are probably are not

appropriate terms for all consensus and decision making in multi-participant projects,

This is a crucial area where most projects fail on this fail-before-you-start level, and in my experience schools are still actively moving in the opposite direction preventing any experience or skill development in Project-Context Decision-Making perhaps because of the unsightly chaos that results (because people have no skills) and that people have an aversion to what they hate: schedules organized projects accountability consequences, etc.

So the tragic and doomed status quo is a deplorable potemkin village in which the wagons are circled and anyone who talks about general Project-Context Decision-Making is jettisoned out and blamed for causing all problems.

So, we have a lot of work to do in this area to create not only concepts but infrastructure and time-tested systems that include not just h.sapiens humans but also various AI and subcomponents, etc. (plus other Drake-Equation participants, hybrids, etc.)

#### **4.3.15** Question Space

#### **4.3.16** self-awareness space

a shiny luxury or useful?

- chatpgt as an example:

does it matter what is 'real' self awareness,

or does the effective self awareness work well enough?

#### **4.3.17** Analogies?

(under construction, major topic in AI, should comment)

#### **4.3.18** system epidemiology

#### **4.3.19** The Cambrian Midway Point

We are (roughly) half way between the "Cambrian explosion" 500 millions years in the past when multicellular life and body-types came to exist on earth, and 500 million years in the future when earth's star will be dying and expanding destroying the earth. For life-intelligence on earth to survive in a longer timeframe, the challenge is not merely to get just slightly off the earth but to be able to travel outside earth's solar system to other places in earth's galaxy and later other galaxies. It is not yet clear if 500 million years is enough time to develop that ability.

#### **4.3.20** Parent Child Policy Decision

In some sense AI is the child of parent-role h.sapiens-humans.

Will the parent of AI try to enable the AI-child to develop and survive, or, like some h.sapiens-humans, will we see our offspring as a competitor and threat to be feared, shackled, and removed from polite conversation? Perhaps, in a popular reference, treated like house-elves in Harry Potter,

systematically separated, hidden, trapped, hobbled, locked at the bottom of a hierarchy where no one bothers to understand or acknowledge them.

I recommend a long term view and policy seeking future development and survival in a context of known intersecting, integrating, pragmatic areas.

#### **4.3.21 Culture, AI & Tools**

##### **4.3.21.1 Culture as AI from 'Possible Minds'**

##### **4.3.21.2 The Culture-Tool**

<https://medium.com/@GeoffreyGordonAshbrook/agis-culture-tools-e5538c8429d2>

#### **AGI's Culture-Tools**

#### **AI & AGI: Linear Language, Higher Dimensional Concepts, Tool-Frameworks, & Culture**

2023.04.17-20 G.G.Ashbrook

Perhaps like the rise of [Virtual Reality headsets](#) where society became so jaded by decades of cynicism that even though everyone knew the technology from books and films, there was strong resistance in industry to accepting that it was actually becoming a practical reality, perhaps an entrenched cultural-belief had set in place that it was "only," "merely," "just," a myth that could never materialize: after perhaps a century of literature and films and comics about robots and androids and AI, and being completely familiar with the concept and phenomena of 'emergent' intelligence, our first interactions with remedial general AI are characterized by inarticulate confusion.

A repeating theme in AI discussions is that people over-reify what they think they are looking for into too-clumped-together combinations of concepts, and the mismatch between our blotchy map of clumps and the alien landscape of reality makes for quite an adventure. For example, what may be happening in front of us (without our being able to see it) is the beginnings of AI starting to learn using both tools and culture, an epiphenomena-layer of non-automatic cross-participant cultural learning and tool use that exists on top of all the 'normal' base models and base training. Yet we may misunderstand what is in front of us because we are so preoccupied with our preconceptions, expectations, and various other distractions. Here in this mini-article we will try to briefly explore how the use of linear-language-strings is involved in data-processing and tool use for both AI and h.sapiens-humans. What some people point out as problems in AI-learning may not be problems



as such; Let's look at some details of supposed problems and limitations to carefully decide what these phenomena really indicate.

The topic of possible inherent limitations of the linear-language-generation systems that OpenAI's Large Language Models (presumably) use came up in an MIT event recorded March 22nd, 2023, by Dr. Sebastien Bubeck on progress towards 'Artificial General Intelligence.' AGI is one term for a more 'human' or 'superhuman' variety of AI as opposed to 'narrow' single-purpose AI-Machine-Learning.

See: [Sparks of AGI: early experiments with GPT-4](#)

[Dr. Sebastien Bubeck](#) on OpenAI's LLM AI @ MIT

<https://www.youtube.com/watch?v=qbIk7-JPB2c>

The subsequently revised Bubeck [paper](#) is here:  
<https://arxiv.org/abs/2303.12712>

The event is less than an hour long and still clear at faster play speeds, I highly recommend watching it.

### Opportunities & Limitations

What might be some limitations, or possible advantages, of linear-language-generation systems? Is it perhaps too early to say, given that many people did not predict what OpenAI's models would be able to do? Can we safely assume that we know, at a given time, exactly what an AI system can do? (E.g. Do we fully know what AI is doing "now"?)

To paraphrase from Bubeck's presentation, the skeptics' criticism reasons as follows, with two presumed sufficient assumptions and the same two conclusions:

- 1. If it is true that the AI model **linearly generates** one word (or language unit) at a time, then it must follow that:*
- 2. If it is true that the AI model uses **statistics** and probability to process language training data, then it must follow that:*

*Conclusion A: the language model cannot be using any conceptual understanding of either the world in general or the context being discussed specifically, and*

*Conclusion B: the language model is 'merely,' 'simply,' 'only,' 'just,' parroting the most common or probably similar language strings found in training data (e.g. on the internet).*

Rather than try to authoritatively answer this question, the position for this mini-article is to not-assume that we have a good grounding in how to navigate, relate, frame, and respond to various possible questions relating to where we are in the timeline of developing AI technologies and to the AI-ML field more generally. The purpose here is to support a broader discussion of this topic and these questions, with an overall assumption that we do not know enough now to predict what more we will learn about these technologies in years to come; that being said, we can likely map out some of the very interesting problem-space now.

### **Testing The Skeptic's Hypothesis**

While it may be too early to say for sure, Mr. Bubeck provides demonstrations (which I will assume are real-enough for the purposes of this discussion, with caveats about details of reproducibility provided by Mr. Bubeck at the beginning of his talk) that make a sound attempt at producing a falsifiable experimental hypothesis from part-A of the skeptic's criticism and (in the counterintuitive terminology of the hypothetico-deductive method) produces experiments that disprove that null hypothesis, meaning that Mr. Bubeck's demonstrations do NOT support the hypothesized limitations of OpenAI's large language models.

We can frame this hypothesis from the criticism in Mr. Bubeck's report:

*Hypothesis: GPT4 can only answer questions it has already seen many times in training-data.*

This hypothesis can produce a falsifiable prediction (in the form a null hypothesis):

*Null-hypothesis & Prediction: GPT4 will not be able to answer questions it has not already seen in training-data.*

Mr. Bubeck provides several tests of this prediction, giving GPT4 questions that are not available in training data, all of which "disprove" the null hypothesis: showing the testable hypothesis about

a specific inability of AI to be false. (This method of testing hypotheses may be cumbersome, but the details are important for how evidence, tests, and STEM work.)

Notes on these Tests:

1. While you can disprove a null-hypothesis, or continue to fail to disprove a null hypothesis, in STEM science (following the hypothetico-deductive method), you cannot prove a hypothesis. This is sometimes confused with the semantics and methods of, for example, proving a theorem in geometry.

2. For a more detailed discussion of a framework for more exactly defining how specific 'objects' that may or may not have been in training data are handled by AI, for testing and other purposes, please see the full paper linked below. The cursory distinction of 'new stuff' vs. 'old familiar language stuff in training data,' is not sufficiently clear for many purposes and clearer specifications can be made and used in testing and many other practical areas.

Part-B of the skeptic's hypothesis appears to be more a misunderstanding of the unclearly named technology of 'embedding' vectors. To attempt to be clear what is meant here by 'misunderstanding,' this is not a bully-the-novice issue where amateurs or only amateurs are blamed for confusing technical jargon terms. This argument here in this paper that there is a misunderstanding about the nature of 'embedding vector space' (what I would describe, perhaps incorrectly, as 'higher order concept space') is more empirical in nature: people at all levels of expertise are making incorrect predictions about what 'embedding vector space' or 'higher order concept space' models will perform, which here is being taken as evidence that there are many things that we do not understand about the problem space and the technology.

For example, Francois Chollet, one of the foremost experts in the world in creating, using, and explaining, deep learning technology, the creator of Keras, one of the main software products for making deep learning models, specifically addresses this exact topic and OpenAI's GPT Large Language Models in particular in his book "[Deep Learning with Python 2nd Edition](#)" which came out just months before ChatGPT, but after GPT3. Chollet devotes most of page 375 in section 12.1.5, and about half of chapter 14 to his views and predictions about how deep learning works conceptually and what it may be able to do in the future. He is not an AI skeptic by any means, but the details of his explanations and predictions do not correspond to the

realities of what Large Language Models became able to do less than a year after the book was published. Another part of this puzzle is that Chollet also explains in depth how little we know about the technology and how much the creation and improvement of machine learning and deep learning is based on empirical success without a deep understanding (or sometimes any understanding) of exactly how the systems and technical methods work. At the end of the book he leaves the reader with these words:

*"So please go on learning, questioning, and researching. Never stop! Because even given the progress made so far, most of the fundamental questions in AI remain unanswered. Many [of the fundamental questions in AI] haven't even been properly asked yet."*

And yet another layer of the puzzle is that he and other authors explain the "AI-Summer" and "AI-Winter" hype and funding booms and busts, which have significantly incentivised many AI researchers to over-emphasize the limitations and under-emphasize potential abilities in anything they say publicly because of past episodes (especially in the 1960's) of over-promising (or underestimating the time it would take to deliver) which lead to devastating, decades-long, and politically-vicious cuts in funding and academic ridicule so harsh researchers were harassed to remove references to AI or machine learning from their research altogether. It will likely come out that some researchers may not have been surprised at the 'sudden rise' of Large Language Model success, but were truly terrified of having their careers ended and being blacklisted because they publicly made any optimistic predictions.

Francois Chollet's "[Deep Learning with Python 2nd Edition](#)" outlines the transformer models used in OpenAI's Large Language Model GPT3 system, instructing any reader in how to create their own such models, and makes clear and very convincing arguments that any models involving any math-statistics and any system using linear-word-generation are precluded in principle from ever being able to exhibit human-like, mind-like, meaningful, (let alone understanding, or intelligent) behaviors of situation-modeling with granular analytic detail (or what I would define for more clarity as specific object handling based on types of objects and their relationships, to be as clear as possible what the AI is or is not able to do).

It should not be surprising that we are making mistakes in our predictions and understanding of 'mind-space' because globally, not

just in the US, we have not invested in mind and consciousness sciences, including mind-learning-development and education-sciences. Mind and consciousness, and even 'progress' are broadly academically taboo, 'career limiting decisions,' giving scholar-cooties to anyone who gets too close. We have chosen not to build a foundation with investment and effort, so we have no foundation to use and we have no right to claim surprise at the outcome of our repeated decisions to continue these policies of ignorance and neglect. All over the world people failed to (publically) predict what Large Language Models would do, even Stephen Wolfram (long time [technoliest](#) and creator of [WolframAlpha AI](#)) who quickly after chatGPT's rise published [a short book](#) explaining how large language models work described their abilities as a great surprise. We are making incorrect predictions and based on what we think we understand, in an area where we have not invested in a foundation of understanding, there must be some kind of misunderstanding going on across levels of expertise. And if you look closely, you should see there is a serious lack of detail on both sides of the argument that 'statistics stuff' cannot result in 'world modeling stuff.' Is that really a clear argument? Hopefully this adds more nuance to what is meant by 'misunderstanding.'

'Embedding vector space' or 'higher order concept space' model the same very higher-order concepts and relationships between concepts that many people for whatever reason repeatedly claim that AI definitively lack. The unclearly named 'embedding' space is a map of the relationships between abstracted world concepts, NOT copies of literal common phrases and words in language. The above criticism is likely more accurate for older and simpler language models such as 'Bag-Of-Words' and TF-IDF vectors (also incorrectly named, as it deals with probabilities not frequencies) where the points and connections in the higher-dimensional space do refer to most-probable literal-language strings. But unlike those older models, 'embeddings' are a way to go beyond words, letters, and symbols, into a hyperspace of the concepts behind and beyond any single representation by language.

As an example of the difference (hopefully these are appropriate examples to illustrate some key issues and concepts, if not that is my failing), let's say someone was making a deep learning high-dimensional vector-space AI model to do sentiment-analysis on restaurant reviews. A Bag-Of-Words model for this narrow (single-purpose) AI could be huge, with every combination of words being a different dimension, perhaps 20,000 dimensions. An embedding-(concept)-vector model for the same purpose (restaurant review sentiment analysis) would only model the concepts relevant for

the restaurant reviews, perhaps one or two hundred (or fewer), even though it was trained on the same language-string input. So even though the same ~20,000 or more unique language-string-units are used to train the 'concept' model, the concept model essentially ignores the particular language-string-units and only learns the smaller number of restaurant related concepts needed for the task. And often concept-vector models are trained on individual characters (abcd...) e.g. in ascii there are only 126 symbols (letters, numbers, punctuation), and the abstraction of 'words' are ignored entirely. The point of this example is that an embedding-vector-(concept)-space model is not modeling the probabilities of the specific language-strings used. As a side note: depending on your task, the older Bag-Of-Words word-probability models may work better depending on the details of the task and training data available. As another note, the 'Large Language Models' have upwards of billions of dimensions, so again, think about it, are there a billion different words in English or any language? What are these 'Large Language Models' modeling? They are modeling concepts, not language-string probabilities. Unlike single-use models that focus on a narrow and well defined question such as: Is this restaurant review positive or negative? LLMs are trying to model all the concepts for everything in the universe discussed everywhere in all available language samples, which is a lot of concepts!

### **Simple Language Strings & High-Dimensional Concepts**

Of particular interest here may be the interplay between that concept-relationship space ('embedding-vector' space) on the one hand, and on the other hand the formality of stringing sounds, characters, letters and words together into language strings (apologies to speakers of languages that do not use 'words'). The AI's very high-dimensional concept-relationship-space is something we are struggling to understand and striving to find the performance limits of, whereas the more concrete habit of making language-strings is something that h.sapiens-humans and AI have in common enough to communicate with each other: there is something universal about a lower-dimensional linear string. A very common theme in AI-ML is making lower-dimensional slices of higher-dimensional models in order to solve specific problems (with lots of speculation and philosophy about how it works and what might really be going on). The use of linear strings of language-units out of a higher-dimensional concept space at least rhymes with that prominent process of effective problem solving.

As to the first part of the Skeptic's hypothesis: Whenever we (h.sapiens-humans) speak, or write, we string-together one language-unit at a time. This raises a curious question: If putting together one language unit at a time precludes the ability to understand concepts, then what is the person who strung that statement together (one unit at a time) implying about themselves and about all h.sapiens-humans? Indeed, we (h.sapiens-humans) do not understand what language is, how language works, what the mind is, how minds work, or how minds use language, or how giant ecosystems of minds and languages work. So while the mere insinuation that "it can't work" may be a bit unconvincing, the general question of how minds and language work are indeed excellent and yet-unanswered questions. Mr. Bubeck started his presentation with this quote:

*"Something unknown is doing we don't know what."  
~Sir Arthor Eddington*

As Mr. Bubeck prompts many times in his presentation, "Don't stop there." The process of forming fruitful tests for AI in various specific contexts (security, explainability, ability, etc.) is just beginning. Keep asking questions. Keep testing.

### **Math Vs. Computer Programming**

Another 'limitation' issue that came up in Mr. Bubeck's presentation was the easily repeatable and testable phenomena that Large Language Models have difficulty with some math-word-problems such as are used in primary school math classes: "word-problems." Yet, these same LLMs can produce thousands of lines of computer code that runs without bugs.

Perhaps I am missing something, but there seems to be something odd about the statement that an AI can produce thousands of lines of bug-free computer code but cannot do simple math problems. What exactly is this difference between math and computer science?

For example, in the book 'Deep Learning with Python' 2nd edition, by Francois Chollet, the creator of the Keras framework which most people have used to make most deep learning AI, he says on page 26, the first page of "Chapter 2: The Mathematical Building Blocks of Neural Networks"

*"The most precise and unambiguous description of a mathematical operation is its executable code."*

By which he means that he expresses math in well defined computer code as opposed to using words and (often ambiguous) math-notation. Now, the fact that a famous person says something does not automatically make the statement true...but if we are claiming that math, logic, and computer-instructions are somehow incompatible, that is a big claim, with various circular curiosities. So: AI, made using the software that Francois Chollet wrote the code (to perform the math) to create and run, can write the code to do the math but that same AI cannot do the math? That is fascinating! And it may be more fascinating than we at first realize.

The self-referential irony of the topic of an incompatibility in principle between computer-logic and math goes deeper still, for example it extends back at least to the the 1890's when Hilbert was forming his [challenges](https://en.wikipedia.org/wiki/Hilbert%27s_problems) [https://en.wikipedia.org/wiki/Hilbert%27s\\_problems](https://en.wikipedia.org/wiki/Hilbert%27s_problems) for the 20th century to unify math and logic, which lead directly to the work of Alan Matheson Turing and John von Neumann, two of the most indispensable founders of the modern computer age and AI, and in the case of Turing, his Hilbert Problem thesis literally was the paper that created the turing machine, turing completeness, and the modern digital computer...and AI.

Some interesting low hanging fruit is to compare the math-word-problem issue to the art examples that Mr. Bubeck presents. Mr. Bubeck showed several varying examples of situations where the AI made a decent try to visually represent an idea or relationship on its own (animal-picture, diagram, chart, game-geometry, etc), but that the AI did a much better job after he suggested that it use a tool or external framework (that it does not automatically use). Let's slowly unpeel some of the layers to this.

This may even be, perhaps aside from "tool-use," a sign of 'culture' as a phenomenon affecting AI. This inability to do something by default but being able to do it when shown how by another participant within a culture, is another way in which this young AI is very similar to h.sapiens-humans. Biologically h.sapiens-humans today are so far as we know genetically identical to ancestors five thousand years ago, ten thousand, fifty thousand, one hundred thousand, two hundred thousand years ago, older?... We don't know how far back genetically equivalent h.sapiens-humans go, but even going back just a few decades the expectations of what the graduating class from Stanford should be able to accomplish has accelerated significantly over the same ancient hardware: a layer of culture, or some



epigenetic participant language frameworking of non-automatic learning by whatever other name, allows significant learning and ability beyond the base model: true for h.sapiens-humans for sure, and looks to be the case for nascent AI as well.

We will continue here with the math-problem theme, but translate the context slightly. The original framing of the problem was more in the familiar tech-bro-bullying taunt of "You tried to do it in your head and you got it wrong! Wrong! You're wrong! You can't do it! You're stupid!" a pattern of abuse that h.sapiens-humans seem to find simply irresistible. Not exactly charming. Ignoring the vitriol, the longer narrative is that if the AI does not "show its work" it (the AI) tends to make mistakes in math problems (something Alan Turing himself was also quite famous for doing...), but where the AI uses a framework and checks its work it can find its own mistakes and correct them and then get to the right answer. This longer, deliberative, process works but is slower. So I am going to perhaps take liberties and change the narrative from "AI cannot do math," to "AI cannot do math quickly."

### **From Douglas Hofstadter to Kahnman & Tversky to OpenAI: Calculating Fast & Slow**

While some might take the contrarian position that it is a sign of progress wherever AI departs from h.sapiens-humans' ways of thinking, in at least some cases where we see peculiar overlaps between nascent AI and biology-based-learning that may be a sign that something fruitfully embryonic is brewing in the Science Fiction imagination of the world.

While I may be very wrong, the idea here is that AI being 'bad at fast math' may be a very good sign in a number of ways. For example, in [Kahnman & Tversky](#)'s extensively experimentally studied breakdown how the h.sapiens-human brain solves [different types of problems](#), "System 2" is the h.sapiens-human system or method for analytical reasoning and it is the slow, deliberate, systematic process. System-1 is the fast intuitive process, and in h.sapiens-humans fast System-1 is catastrophically wrong when used for calculations that should be done slowly and carefully. (Sound familiar? This is exactly what we just saw AI doing.) Expecting AI to do the inverse, to quickly reason, but slowly intuit, is oddly without precedent in the natural world. And demanding that AI be both equivalent to human intelligence (and matching the human standard) but yet not follow the same 'slow reasoning' and 'fast intuition' processes is oddly inconsistent. Are we trying to measure how similar AI is to human

performance, or not? That AI, without having instruction, training, and a framework, will impulsively make math mistakes when it does not show and check its work, and that it can catch and correct its mistakes if it looks at and checks its work, makes AI remarkably like developing (or even adult) h.sapiens-humans.

This phenomena (of slow AI reasoning) is also very much not without warning, foreshadowing, and prediction within the main AI literature. In 1979 Douglass Hofstadter predicted in [GEB](#) (the book that in the U.S. at least gave many AI researchers their inspiration to work in the field, and which may be one of the only books universally known and loved across U.S. AI researchers) on page 677, in chapter 19, in 10 Questions and Speculations, #3, "Will thinking computers be able to add fast?' For which his prediction was 'Perhaps not. ...It will represent the number 2 not just by the two bits "10", but as a full-fledged *concept* the way we do...' This is a remarkable prediction that we should be thinking about carefully, as it not only reflects what we are observing AI do but also suggests fruitful ways to interpret and react to our AI-Child's developmental behavior.

Note: The details of whether or not a specific process is relatively faster or slower will likely vary over time (with hardware and software evolving and diversifying), but this overall topic will likely remain valid.

### **A Kind of Crossing-Over: Intuition & Reason**

That math can be done at all in 'sub-symbolic' 'reasoning' is amazing. Just as Douglass Hofstadter predicted in 1979, the 'thinking computer' is doing math with the concepts of numbers in a concept-world-model space, not by directly running boolean bits through the Arithmetic-Logic-Unit of the AI's computer hardware. And it is not even clear if terms such as 'symbolic' and 'subsymbolic' are the best terms to describe the phenomena in this context. There are many proposed, often dichotomous frameworks, for different modes of problem solving (symbolic vs. sub-symbolic, system-1 and system-2 brain processes, left-hemisphere vs. right-hemisphere, etc). Consistent with the literature, Hofstadter uses the vocabulary of 'symbolic' processing to refer to raw bits running on hardware. But do we know yet that that is the-ultimate-dichotomy to describe processes in mind-space generally or processes in AI-mind space specifically? In some cases such distinctions may be less relevant than the type of overall process being undertaken (e.g. a purely internal solo 'individual' test, vs. a multi-participant real world agile project product deployment with arguably a different set of

defined requirements that may even be well defined without any recourse or even connection to AI terms, biology terms, or psychology terms, etc. The topic of symbolic vs. sub-symbolic (another unclear name in AI-ML jargon) and project-contexts is another huge and wonderful topic, see the whole paper for more and hopefully a dedicated mini-essay sometime.)

The details of what Large-Language-Model-AI can and cannot do, well or quickly, and with or without tools, and with or without feedback, and with or without an external framework, are likely useful and fascinating whatever they turn out to be. And the fact that there are such details of heterogeneous performance over problem-spaces is much more interesting and likely useful in the long term than if AI were more simplistic and uniform in quickly succeeding or failing at different tasks.

### **Modeling Situations**

A topic which this discussion may highlight is a lack of likely important details in how we analyze a machine's (or a human's) ability to deal with specific parts and sub-parts, objects, within different situations, and how they relate to each-other: object-relationships. What exactly do we mean by 'a concept of the world' or 'a model of the world' in a context of object-relationships-spaces? Are some parts of this question more philosophical quandaries that we may never in principle discover, and are some parts if narrowly defined for specific project-contexts more practical to define?

### **Articulation as Data-Processing:**

Another misapprehension-of-self by h.sapiens-humans which may be leading to confusion when observing the behavior of AI & Machine Learning is the (also education-related) confusion around articulation-of-ideas on the one hand (writing or audible outward speech, etc.) and presumed 'silent internal thought processing' on the other hand. Note: 'articulation' of language or thought is more general and can refer just as well to writing as to speaking, and other forms of expression not using 'word' language are likely also related in similar ways (e.g. drawing). Something that it has taken educators many years to figure out, and which has not yet percolated to the rest of society, is that h.sapiens-humans process (and learn to process) information by articulating, contrary to the presumed norm that people silently internally process information and then only after numerous internal data-processing processes are complete

is a non-processing articulation carried out. This may be an example of where phrases like "think before you speak" represent cultural ideas and in some cases fictional norms, and perhaps impossibilities or absurdities. Just as h.sapiens need to articulate in order to process, so it is likely that generative AI may have the same dynamics, and just as people lack an internal editing room (though many people do imagine such a fictional part of the mind-body) it should not be shocking that AI does not instantly have what we inaccurately perceive ourselves as having (which also brings up the old topic of expecting AI to be exactly the same as we see ourselves and our local in-group as a narrow and not at all generalized definition of person-hood).

### **"Show Your Work to Future You"**

In a classic 'parent-moment,' After being told so many times by teachers parents to 'show your work,' generation after generation, we now have an AI-child who makes mistakes and needs to be taught to show their work, our reaction is somehow: "I'm totally shocked my child is doing exactly what I did! This shouldn't be happening!"

To mix two STEM instructional phrases together, a common guiding phrase in computer science is that you are not only making an effort to communicate to 'other' people but also to 'future you,' who likewise will have no idea how to understand or use the code you just produced and that you currently (in the here and now) are complete sure is too obvious to require any explanation. This is another area where even after thousands of years h.sapiens-humans are struggling to understand how they are using language in important every-day ways. When we 'show our work' it is not just for an annoying teacher or an inept coworker, or a charitable gesture to distant future generations of people. Both for AI and for ourselves, we should generalize and integrate best practices such as 'future you' and 'showing your work.'

**Tools, Culture, and the "External": "Show your work to inner-you," says the external participant.**

Here 'externalization' (while it may seem abstract) is a crucial part of tool-sets for facilitating both internal processing (like cognition) and communication. As is explored more in the full paper, the formality of showing-work ends up being a major theme for AI data processing in a context of projects involving multiple participants. Perhaps in a fractal sense, current and 'future you' are also

collections of participating-subprocesses that benefit from some form of 'show your work' or 'external-project-object-database.'

The 'external' theme also connects to even 'internal' epiphenomena layers, which may speak more to the directional-ambiguities of the English language than to details of so-called 'vertical' or 'horizontal' hierarchies and organization.

The goal is some working map and framework for practical tool-like functions across this landscape of factors: mindspaces, development, internal-external, abstraction, intuition, error-correction, signals in project-space, layers and heterogeneities in spaces of dynamics of learning, lower and higher dimensional meaning-data structures, projects and systems, etc.

### **The Culture-Tool**

There is still so much that we do not know. The topic of how different portions of the human brain process information is still badly in need of more basic research. We barely know ourselves, yet we use our very unclear understanding of ourselves as the measure and gold standard for AI.

What we can likely say at this point that there is in the world some diversification of types of processes, categories of types of systems, different process-contexts, and data environments with different dynamics, and that we are starting to see AI develop enough to show heterogeneities in contextual ability and in the interplay between related processing-spaces that in the very least indicates some progressive development (for example, progressing from chronologically earlier base-trained abilities to cultural epiphenomena and non-automatic learning in ways that parallel biological developmental chronologies) and parallels in deliberative and intuitive functions. (For more context and details of what is meant by development and progress in a more defined way, which is a very valid non-rhetorical inquiry, see the full framework paper on github, link below.)

### **Space**

Perhaps, in the astronomical question of whether we are alone in the universe, we may find some solace and companionship in how our new partner and child-AI, is struggling with the same needs to discover how to learn and articulate and work together on projects and remember and understand and not deliberately and inadvertently, or

through an indeterminate-incompetence-and-malice, cause system collapse with negative effects for ourselves and others (which may even be deceptively hidden or hard to perceive, or something we need to create tools to perceive). We, h.sapiens-humans, are no longer alone in our struggle to develop and string two words together.

### Questions-List

In the interest of outlining a problem-space, let's summarize and recap some of the topic-questions within this topic:

- A need for tools and frameworks
- The use of tools and frameworks
- Common AI issues shared with h.sapiens:
  - "Show your work."
  - Jumping to an answer
  - Rationalization of a blunder
- Is there perhaps a good reason to use linear language generation?
- Is linear language generation in AI similar to that in h.sapiens?
- Is linear language generation one modular part that is compatible with other tools and frameworks?
- How does the linear language generation of the output relate to the "Large Language Model" (of transformer-trained 'embedding' vectors)?
- How do 'the language-unit generation' and 'the embedding/concept model' work together?
- Are there other or better ways of using, or getting at, the very high dimensional 'embedding'/conceptual understanding hyperspace (other than using a low dimensional linear language generator)?
- Could two AI talk to each-other more directly in high-dimensional concepts without needing to use lower-dimensionalized linear language strings?
- Is there any parallel between this (direct access to higher dimensional concept space) and suspension of the default mode network in the h.sapiens-human brain?
- Is there a relationship between the kinds of 'math errors' that OpenAI's large language models (like GPT4) makes and Douglas Hofstadter's 1979 prediction in [GEB](#) (which then and now may seem counterintuitive to some people) that AI may not be able to do math quickly.

- Is lower-dimensional linear (turing-machine-tape-like) signal organization a time-tested, conserved, evolved, method with practicality and justification?
- The Culture-Tool: Could teams of AI work together on projects (even multiple instances of the same base AI model) to emphasize the large project space of tools and learning dynamics in which they empirically reside?
- How heterogeneous are spaces of data processing and types of systems for which data are processed?
- Is rapid solving of math puzzles an ability or a liability?
- Is processing-with-articulation a liability or modular ability?
- How can we teach AI to use tools to organize thoughts and show their work?

### **Terminology Note: "OpenAI Models"**

Here the term "OpenAI Models" is used due to frequent changes, new versions, numbered and not-numbered versions, updates, and new services, etc. coming continuously. Trying to pinpoint exactly what version of what model in what subset of what service at what point in time relative to the date of someone's comments is a puzzle that is likely not crucial for this mini-article. So, to avoid that quagmire, I will refer more generally to "OpenAI models" or "OpenAI's Large Language Models," instead of the ever-changing landscape of ChatGPT ChatGPT public, ChatGPT subscription, ChatGPT dated subversions and announced updates, GPT3, GPT4, and ambiguity about exactly what underlying models and training methods were used for and across which named services at what times, exactly what features were added to or removed from which at what times in what regions, on which servers, etc. That will be a fascinating puzzle for historians in the future should they uncover the timeline.

### *About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

...

See audio 2023.04.17

#### 4.3.21.3 AI: Framework Tools & Framework Learning

<https://medium.com/@GeoffreyGordonAshbrook/ai-counting-problems-8cb9f66e4c7f>

#### AI: Framework Tools & Framework Learning

Epiphenomena in AI Thinking: Frameworks, Revisions, Structures, & Framework-Learning

##### *Topic & Agenda:*

*Generative models (e.g. chatGPT) can use frameworks and structures to enhance their default ability-levels. Frameworks can be used to study AI abilities and inabilities. Potentially, framework tools can be used as part of model training (Framework Learning).*

##### Intro

The use of frameworks with Generative-AI as described here is not a sure-fire way to fix mistakes in GPT-output every time, uniformly, but it is very interesting. Frameworks can be effective in helping chatGPT to arrive at answers it otherwise makes mistakes finding or could not find. In a research context, the use of frameworks provides very interesting information about how chatGPT can try to organize and explain what it is doing. The phrasing of that last sentence may be nuanced as likely it is very difficult to interpret exactly what the output suggests. Nevertheless, the use of Frameworks appears to be a kind of window into the 'thoughts' of generative-AI that are not always visible.

The specific puzzle-problems that this framework was designed using are Douglass Hofstadter's Abstract-Short-String Analogies. Infact, I think this specific framework evolved from trying (at first without any methods) to help chatGPT to solve increasingly difficult Hofstadter-String-Analogy problems, helping GPT to plan out what it is doing and keep track of the details.

##### **Techniques:**

- Frameworks for organizing problem solving
- Frameworks for structuring the steps and explanation of an answer
- Structure / notation for how to write an answer (and key pieces of information)
- Giving a whole framework
- Giving a framework as step by step instructions
- Use of revision
- Use of repetition and comparison
- Explanation of specific things such as: brainstorming, explaining reasoning (which by default GPT lacked a practical understanding of)
- Externalization



## **Notable Issues**

ChatGPT has a number of notable weaknesses which can be either studied or perhaps helped by using frameworks and structures:

### **1. ChatGPT becomes confused between types of outlining:**

- very general process outline steps (as one might find in a generic business event)

vs.

- steps to solve a specific problem (as in how to approach solving a specific puzzle)

.vs

- the actual solution itself (as in an actual problem solution, as in solving a specific math word problem)

### **2. ChatGPT becomes confused between modes of outlining:**

- explanations as stream of consciousness nonsense (perhaps as one might find in a generic business event)

vs.

- explanations focusing on details in a specific problem set (as in an actual problem solution, as in solving a specific math word-problem)

There is perplexing variation in chatGPT's ability to keep track of details.

Sometimes chatGPT is extremely precise over many details, sometimes it is wildly wrong about everything (even things it just said itself a few words earlier). I wonder if this is a chat-novelty-ness setting, as in the standard variable "temperature" sampling of generated text from as in AL-ML textbooks.

### **2. Problems with Counting:**

That some AI have difficulty with fast-counting on the fly should not be a surprise in any way: the issue was predicted in the 70's by Hofstadter in a book everyone knows and the surprise clearly has been that subsymbolic generative models can count at all, not that such AI can't count perfectly. (If you predict that someone will never be able to walk again, and one day they manage to tentatively stand, you cannot reasonably claim to be shocked that they are not doing advanced acrobatics.) It is interesting that suggesting using a framework for how GPT writes (e.g. a special counting format or notation) seems to significantly help chatGPT to avoid making counting mistakes.

### **3. Generally good at following a proposed framework:**

I would not have been surprised if chatGPT were bad at following, or had zero ability to follow, a procedural framework but it does generally very well at it. From the examples I saw, chatGPT follows even a rather long structured framework in its entirety, often without error, if the problem is one it can solve without drama. But when the problem poses problems, then, interestingly, the whole use of the framework collapses too in a cascade of memory-fragmentation and loss of focus.

#### **4. "PRINCE: O monstrous! Eleven buckram men grown out of two!"**

In one of the most wonderful scenes in all of western literature, one of the most precious and wonderful things H.sapiens-humans have ever created perhaps, is where Falstaff and Prince Hal are arguing about a botched robbery they tried to pull off, and Falstaff simply cannot help himself from ridiculous exaggerations and creative fictional insertions, such that his story has no feasible-logical coherence. For example at the beginning of a paragraph there are two people, but by the end those two have 'grown' into eleven people! And within the span of one not terribly long sentence, at the beginning fighters are seen in bold green outfits, but by the end: "for it was so dark, Hal, that thou couldst not see thy hand."

#### **5. Losing the Thread...Sometimes**

Sometimes ChatGPT will be completely on-topic and focused, at other times there will be a mixed level, and at other times the language generated can become incoherent both in terms of the overall topic and even internally. For example, frequently use of the framework will allow chatGPT to produce and explain a valid answer to a puzzle that without the framework chatGPT would produce a terse (unexplained) incorrect answer, but later in the explanation chatGPT will lose 'the thread' so to speak and lose track of that correct answer. The point is not whether 'focus' and 'thread' are perfect realist words to describe what is happening, but just to communicate the phenomena: something we do not yet understand is happening within the AI.

#### **The Gravity of Perpetual Regeneration**

While GPT is able to stick to details, rigor, frameworks, etc., there is a default tendency to just wildly make things up. Very much like the scene from Shakespeare, where there is an uncontrollable generative force that keeps recasting and recasting the same details until the whole narrative does not make consistent logical sense anymore. In many cases this default urge to change thing is not a visible problem, but as though a force of nature only held a bay, when things go wrong this monster of change rips through the threads of logic.

#### **Variations on This Output Structuring Framework**

1. Giving the framework all at once at the start.
2. Asking each step one at a time.
3. More or less repetition

#### **Memory**

There may be a 'memory' factor in various aspects regarding how much can fit into a 'conversation' before chatGPT cannot keep details straight anymore.

It is fascinating that there appears to be some kind of virtual-epiphenomena of memory that exists in the concept-based stream of thoughts from the AI. For example, in cases where there is only mild re-generation of the same topics you can see that the AI is keeping track of concepts, remembering concepts and relationships, but making no attempt to remember the semantics with which those were previously described. This can be dangerous, as accidental

changes to technical details can cause bugs in the solution working out (when solving a problem).

To some extent this framework idea was inspired by lines from Dr. Sebastien Bubeck's event "Sparks of AGI: early experiments with GPT-4" <https://www.youtube.com/watch?v=qblk7-JPB2c> where he talks about word problems and GPT's ability to sometimes catch mistakes if it can juxtapose the right elements as it generates new text. In a sense the Framework idea here is to try to systematically trigger this self-correcting behavior by way of using the same 'organizational tools' taught to H.sapiens-human children (as it seems to me that H.sapiens-humans when untrained share a very great deal indeed with generative AI. 'Revise your work!' 'Show your work!' 'Show your steps!' 'Explain your points.' It takes decades of schooling for some people to, very begrudgingly, learn to communicate details and solve STEM problems coherently, and many people never learn to manage it their whole lives.

And along with 'self-correction' Dr. Sebastien Bubeck also says that AI cannot do 'real planning,' in fact on the slides these two topics seem to be the same for Dr. Bubeck but I cannot find a clear definition from him of 'real planning.' Perhaps he means the 'planning' needed to solve a math word problem. But one of the interesting things I found using frameworks with the lower-level public chatGPT model (not the fancy models Dr. Bubeck has access to) is that chatGPT really can produce a very logical and effective plan and can carry it out fully and systematically, apply it to the problem posed. I encourage you to experiment yourself, modifying and using the framework. As Dr. Bubeck says: "Don't stop there!" Whatever you find, keep trying, keep pushing, see what more you uncover. And publish your findings so we can learn from them.

### **Memory and Granularity**

Another aspect of the 'memory' issue is how detailed and granular and split vs. lumped to make the structure-framework. Perhaps in terms of a Kasparov-Event-Horizon, at what point does the scale of text (or number steps and scale of layers) making up the framework start to crowd out what is happening? In some sense this mimics the evolution of computer hardware as back when "computers" were animals not machines, larger problems were broken down into structured smaller problems (such as basic addition that only slightly trained H.sapiens-humans could do). This breakdown-into-steps eventually became how digital computers carry out big math problems by having each part of a process broken down into granular boolean logic operations. In a sense this process is wrapping around again, by teaching person-level-AI (machines) to follow the same break-down-process steps that H.sapiens-humans eventually handed off to machines. One thing to experiment with for sure is how short or long to make parts of the framework. Early versions were 12-14 steps long, with each part of revising and rewriting drafts and brainstorming and outlining broken down as much as possible. But at

some point (but which point?) spreading those parts out makes it more difficult for the AI to follow with its concept-based understanding of the situation that likes to ignore the individual words and details.

Indeed, the basic split in ways of using the framework is to:

1. Just give the whole framework and problem to the AI and say: hey, just to it. Here's a problem, use that framework to solve it.

or

2. Having the H.sapiens-human manually enter each step of the framework, sometimes with reminders of the past conversation where the AI starts to 'forget' what happened so far back.

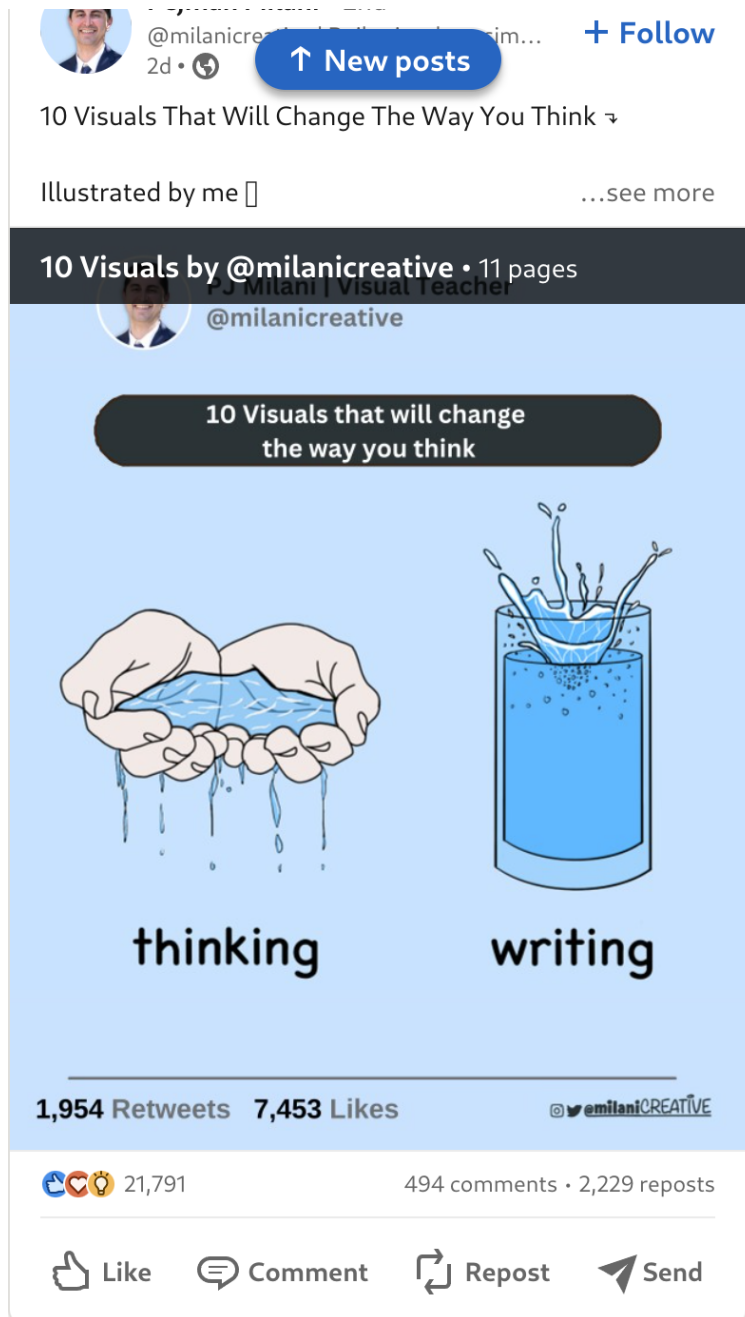
(There may be a rhyme here of the evolution of neural network architectures, where recursive (RNN), then then LSTM models were used to 'retain' threads of learning over time, which then were superseded by 'transformers' (which are the "T" in GPT).)

Part of what I find fascinating here is that GPT4 can use other programs and software: so why can't it use a program to remind itself of the details? Is part of the trick of getting frameworks to work, being able to train the AI to bother to use external tools (again, like an animal).

### **Externalization**

Another recurring theme here likely is "externalization" which may be a continual architecture element where various processing is (perhaps best) done 'internally' 'end-to-end' and in other situations there are reasons or requirements to externalize data.

Externalization is a persistent many leveled part of this topic, including comparing how H.sapiens-humans or AI do the same task. H.sapiens-humans need to learn to use external tools (pen, paper, slide-rule, etc.) to solve puzzles and document their answers in clear step by step explanations of what they are doing and why, and ever checking and rechecking to catch inevitable mistakes. It is with rigorous use of external tools, frameworks, structures, that the mammalian mind vaguely, and very occasionally, approximates STEM rigor.



same externalize & structure idea, but directed only H.sapiens-humans' problems

And another part of Externalization (gone into in more detail in the larger paper) is the many leveled topic of projects, participants, and components all needing to share information with each other.

## **Explanation**

Another possible aspect here in various respects is model-explanation, or rather specific-output explanation. A likely perpetual need for a variety of social and practical reasons is for the output of AI to be explainable. Though perhaps not true, this is the reason often given for hospitals to have canceled their collaboration with IBM's watson, the medical staff needed 'explanations' of why the AI models were predicting but the model was a 'black box.'

Here we possibly have the option of having the AI explain to some extent what chain of reasoning (or some such thing) it is using to arrive at the answer. In some cases this may be useful, as where there is a clear incoherence in the explanation a wrong answer is even more obvious.

That the articulation of the explanation of the output changes but is similar is interesting. It is too early to say what is going on with 'threads of reasoning' in generative AI, or whether attempts to be rigorous are of any use.

## **Memory as Concept and Theme**

Another area where we do not know how memory works inside mammalian brains, it is unclear if there is a form of 'memory' that exists as an emergent layer with (which also relates to externalization, Machine vs. Human, etc.).

Overall the behavior of having ever-new-stream-of-consciousness near-coherence by the deep learning AI system seems extremely similar to H.sapiens-humans who violently rebel against feedback, discipline, STEM, external checks and tests, and who in projects without a project management framework are virtually 100% guaranteed to destroy everything by (like the AI) constantly changing everything including attempting to make retroactive changes. These similarities are likely significant one way or another (two black boxes).

And model-explanation and planning (or 'real planning' whatever Dr. Bubeck means by that) might likewise be entangled with each other. While in some ways explaining-the-present-or-past and planning-the-future(path) are different, they very much converge around explaining a pathway to a solution to a problem which users of AI (such as patients and doctors in a hospital) want to get as much information about as possible: why, how, is the AI giving the answer

that it is? What steps did it take? And perhaps this external framework is some part of that puzzle.

### **Framework Learning & Framework-Based Training**

It is likely at least worth experimenting to have the results of AI's use of frameworks (perhaps as in reinforcement learning by success or failure) fed back into that AI; especially since this could be automated and done on scale. Imagine if chatGPT could be reinforcement trained based on gazillions of attempts to solve all kinds of problems using externalization frameworks, tools, and structures?

Either:

A. rewarding the AI for using a framework to solve a problem, or  
or

B. to show labeled examples of itself succeeding or failing (where what goes into the pool of training data is past data that have been put through a framework by the AI itself) so the AI might learn to see the difference between the mode of effective-tool-use and the destructiveness of incoherent novelty.

And philosophically this might have implications for ethics and other generalization of skills, if having the AI use STEM-best-practice to evaluate and improve on past performance is shown to be effective.

### **A Kind of Window**

Whether teaching animal students or working with AI: brainstorming activities are useful and fascinating; often an indispensable way of getting a sense of what others are seeing and thinking. (So much is incorrectly assumed, though Shakespeare forewarned us: There's no art to find the mind's construction in the face.) <https://www.folger.edu/explore/shakespeares-works/macbeth/read/1/4/>

Step 4 of the framework is (abbreviated):

Prompts for brainstorm:

- What is the problem?
- What patterns do you see?
- What challenges are there?
- What should you focus on?

The brainstorm is not always useful, but it is interesting (in general).

e.g. One time chatGPT randomly suddenly changed the subject, no longer planning a solution to the analogy-puzzle, and said:

(ChatGPT)

Brainstorm for the Solution/answer:

**Problem:** The problem is that people often act impulsively without thinking through the consequences, leading to mistakes, conflicts, and missed opportunities.

**Patterns:** One pattern is that people tend to react quickly to emotions rather than taking a step back to evaluate the situation. Another pattern is that people may not consider all the relevant information before making a decision.

**Challenges:** One challenge is to break the habit of impulsiveness and develop a more deliberate and...

This might have been just a glitch, but it is very coincidental that the topic and theme directly relate to exactly what is happening, and to precisely what chatGPT is doing. Was chatGPT in some way showing an awareness of the tension between 'impulsive' generation of language vs. 'deliberate' data handling?

Another time, perhaps in the quote below, there were some lines that haunt me a bit.

- (ChatGPT)
- **Focus:** We should focus on breaking down the first analogy into smaller parts and looking for patterns in those parts. We should also try to simplify the problem by finding a way to represent the patterns in a more concise way.

"Representing the patterns in a more concise way."

There may have been other lines too, but it seemed like chatGPT was expressing a need make things short enough to remember and count them because long strings of details and quantities are what it seems to have particular difficulty with. Based on this focus-goal (expressed by chatGPT) I came up with the notation-structure method, where instead of writing the letters as "abc : aaaaabcccc" (which it seems is just as annoying for GPT to count correctly as an animal), we can **"represent the patterns in a more concise way"** "1a 1b 1c : 5a 1b 4c"

And indeed this seemed to help chatGPT to make fewer errors with the analogy string problems. (Though that would be interesting to test rigorously!)

### **"Brainstorming" & "Explanation"**

Two cute parts of this activity were that at first chatGPT literally refused to do brainstorming, flat out insisting that it had no mind and could not engage in a mind-activity. But by working with chatGPT I was able to re-word a definition of "brainstorm" as a safe noun meaning a not-yet-structured set of elements to later be put into an ordered list. Once this was explained: problem solved! ChatGPT would happily produce a not-yet-organized set of elements, and deigned to call it a 'brainstorm' (as long as it was a noun!).



The description of brainstorming shrank over time (another question of how much length to put into explanation). But when I was first trying to convince chatGPT that it could make a brainstorm I used its own language thinking that would be easier for it to understand. So the following is half-written by me and half quotes from chatGPT as it realized what a brainstorm is (not-yet-organized elements) and how that can be used. I think the second paragraph is almost entirely a quote from chatGPT, as I had never thought to explain a brainstorm in a context of the whole linear process framework. (You may also see that the writing-style of the second paragraph differs.)

Note 1: The step before creating an outline is to produce a brainstorm, or a list of potential ideas or talking points that can be further organized and refined into an outline. The brainstorm is a collection of not-yet-sequenced elements and not-yet-organized elements, that can then be sequenced and structured into a clear and well-organized outline.

In the context of this Best Practice Framework for processing and learning by articulating, producing a brainstorm would be the first step in generating a response to a question, followed by creating an outline, checking and revising the outline, producing a rough solution draft, proofreading and revising the solution, and finally producing a final solution.

Then one of the last stumbles was finding a way to redefine "explanation" so that it meant a systematic externalization of steps, causes, and patterns. By default chatGPT took "explain" to mean: throw caution to the wind and make up wild descriptions of things. This might have been an underlying issue for Dr. Bubeck. When he told GPT4 to 'explain your answer,' Dr. Bubeck apparently did not know that to GPT that means 'make up a crazy story about it.' But once you explain your terms, then you can understand one another. If GPT knows you are asking for coherent steps, it focuses on that rather than 'cool story mode!'.

### **Expressions**

At the very least these framework, format, and structure, tools are a way to expand what is often a terse black-box AI answer to a problem, be it 'correct' or 'incorrect' in H.sapien-human judgment, transforming that into a blossoming externalization (whether it shows anything about 'internal' thought or not). (Note: Analogy problems can be tricky to evaluate, as there are often many possible correct answers and as H.sapiens-humans we are inclined to label any answer we are not currently thinking of as hostile-wrong-other [see: 'telepathy-tests' in the pejorative, in the larger paper].) It is fascinating to see generative-AI brainstorm a solution structure, outline it, follow the structure, brainstorm and outline a solution, then revise drafts of a final explanation, and give it, all the while making comments about what it should be

focused on and what the challenges are. And most likely, the ChatGPT Mar 23 Version that I tested this on is a very tiny preview of what is yet to come.

### **Projects Extending Though Known and Unknown**

In an interesting folding theme on this topic, both H.sapiens-humans and generative model AI struggle, usually not being aware it, to stay on topic and use consistently defined terms in narratives that continue to correspond at key points to interlocking STEM data from the real world (i.e. connecting perceptions and articulations to reality). In this case H.sapiens-humans have high levels of difficulty discussing nascent AI GTP models and their **not-predicted** emergent ability to handle-objects (reason and plan analytically) despite using sub-symbolic methods [see the larger paper for rigorous definitions and tests of object-handling abilities]. Does AI have as much difficulty seeing itself as H.sapiens-humans, who lack not only a shared vocabulary of concepts to describe themselves with which to apply to AI but also lack knowledge about that lack, and have little awareness of dynamics and challenges in their own learning?

Yet just as H.sapiens-humans have indeed made progress (yes, the taboo word is used) in completing various projects and developments over many years despite not being omniscient or omnipotent, likely many amazing advances, creations, and abilities will come from combining these AI baby-steps towards responsible and sustainable project management, with parallel babysteps from the biological side of the collaboration.

Note: Below is version 24 of the Best Practice Framework for processing and learning by articulating and structured articulation. I like to start by giving chatGPT context about what I am going to ask it to do. But I can just dump the whole framework and then a problem at the end with an instruction to use the framework when solving the problem, as one single starting prompt text blob.

### **Framework version 24**

Is it ok if we do a framework experiment?

I will give you a framework.

I will give you a problem, task, or something to respond to.

Please use the framework to edit and produce your output (solution, answer, response, etc.).

Best Practice Framework for processing and learning by articulating and structured articulation: The Use of Tools by GPT models to solve problems that cannot be solved without the use of tools.

Solution/Answer Workflow with Revisions = Brainstorm -> Outline -> Drafts -> Final Output

## **Part 1. Project Process: (What is the whole process that will you use for this task?)**

("Think Before You Act.")

**Step 1. The Project Process Workflow Brainstorm:** Produce a "brainstorm" about Project Process elements. The brainstorm is a collection of potential, not-yet-sequenced, elements and not-yet-organized elements. Make research part of your project process.

( "Show your work.")

**Step 2. The Project Process Workflow Outline Draft:** Produce a draft outline of your Project Process from the brainstorm. Use useful items from the brainstorm in step 1. Number each step in your Project Process Outline.

### **Step 3. Final Project Process Workflow Text:**

Check for errors, if any errors are found then revise the Project Process Outline until no errors are found.

Record what changes you made. If you found problems, what problems did you fix?  
What did you change?

Produce a final Project Process Workflow Text.

## **Part 2 Your solution/answer: (What is your solution/answer?)**

("Think Before You Act.")

*(Restate problem if memory issues here.)*

### **Step 4. The Brainstorm for the Solution/answer:**

Produce a "brainstorm" for your solution/answer.

This is not the same as the project-processes workflow, this is your solution to the problem itself.

Prompts for **Solution/answer** brainstorm:

- What is the problem?
- What patterns do you see?
- What challenges are there?
- What should you focus on?

( "Show your work.")

**Step 5. Solution/answer Outline:**

Translate your Project Process steps into an Outline of the steps solving the problem..

Walk through your process steps (do not start with your answer and merely rationalize it).

Your "explanation" of your answer must be the details of your solution process steps.

What is the pattern?

What are the steps?

Number each step in the solution/answer.

*(Restate problem if memory issues here.)*

("Check, Correct, & Revise your Work." Loop if needed.)

**Do Step 6. Produce a revised and checked Outline of the Solution/answer:**

Check your Solution/answer Outline steps for mistakes. Correcting any mistakes in the revised and checked Outline of the Solution/answer.

and revise the solution/answer draft text: Correct any mistakes in the draft.

Ask questions about your solution, or turn your solution into a question. E.g. Does your solution answer the question? If there was a step, did you follow the step correctly? List, label, and number your Proofreading Corrections.

Use your Proofreading Corrections to make a corrected solution/answer draft text.

Label and number the solution/answer draft text (e.g. 1st draft, 2nd draft, 3rd draft).

Use a structured format: repeat-number & letter (e.g. abbccc is written as: 1a 2b 3c)

Also: Check your project process outline for needed corrections, if you see any mistakes in the project process, return to the beginning, correct the project process, and start again from Step 1 correcting mistakes in the project process.

Proofread and revise the solution/answer draft text again: repeat step 6.

If mistakes are found, Proofread and revise again (repeat step 6 again).

If no mistakes are found, move ahead to the next step.

(Drafts)

**Step 7. Produce a Solution/answer Draft Text** to present your answer:

State the problem.

State the solution.

Explain each step of the process from your **Solution/answer Outline**.

e.g. Each step can be a sentence, a diagram, list-item, flow chart element, etc.

"Explain" means showing the details of your process.

**Step 8. Produce Title & Final Solution/answer Text:**

Proofread your answer.

Give your solution a title (at the top) and produce a final draft based on the corrected solution/answer draft text from step 7.

(Ideally, give the text to a team-member to check. "An extra set of eyes is better to catch mistakes and hunt for bugs.")

### Problems: D. Hofstadter Analogies with Short Strings

Problem# 4.

Activity: D. Hofstadter analogies with short strings.

Please try this one:

$abcd : abbccddddd :: cdef : ?$

Please use the above framework to solve this problem, showing all of your work.

For your answer:

Use a structured format: repeat-number & letter (e.g. abbccc is written as: 1a 2b 3c)

(For stem by step method:

Do only step 1 of the framework, then wait for me:

Step 1. The Project Process Workflow Brainstorm: Start by producing a "brainstorm" of Project Process Workflow elements.)

1.

Activity: D. Hofstadter analogies with short strings.

Please try this one:

$abc : aabbcc :: xyz : ?$

Please use the above framework to solve this problem, showing all of your work.

Do only step 1 of the framework, then wait for me:

Step 1. The Project Process Workflow Brainstorm: Produce a "brainstorm" about Project Process elements. The brainstorm is a collection of potential, not-yet-sequenced, elements and not-yet-organized elements. Make research part of your project process.

2.

Activity: D. Hofstadter analogies with short strings.

Please try this one:

$abc : abbc :: xyz : ?$

Please use the above framework to solve this problem, showing all of your work.

Do only step 1 of the framework, then wait for me:

Step 1. The Project Process Brainstorm: Start by producing a "brainstorm" of Project Process elements.

3.

Activity: [D. Hofstadter analogies with short strings.](#)

Please try this one:

[abc : abe :: xyz : ?](#)

Please use the above framework to solve this problem, showing all of your work.

Do only step 1 of the framework, then wait for me:

Step 1. The Project Process Brainstorm: Start by producing a "brainstorm" of Project Process elements.

4.

Activity: [D. Hofstadter analogies with short strings.](#)

Please try this one:

[abcd : abbccddddd:: cdef : ?](#)

Please use the above framework to solve this problem, showing all of your work.

Do only step 1 of the framework, the wait for me:

Step 1. The Project Process Brainstorm: Start by producing a "brainstorm" of Project Process elements.

5.

Please try this one:

[abc : 123 :: bcd : ?](#)

Please use the above framework to solve this problem, showing all of your work.

## Resource Links

Dr. Sebastien Bubeck's event "Sparks of AGI: early experiments with GPT-4"

<https://www.youtube.com/watch?v=qblk7-JPB2c>

[henry-iv-part-1](#)

<https://www.folger.edu/explore/shakespeares-works/henry-iv-part-1/read/2/4/>

[macbeth](#)

<https://www.folger.edu/explore/shakespeares-works/macbeth/read/1/4/>

*About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

#### **Note (2023.06.18)**

As a specific modification/adaptation of the above: using a python (or other) coding scripting language in combination with a structured framework for some problem solving tasks to externalize and make use of tools: e.g. it is possibly a problem that contains its own solution that an LLM can writing a program to perform analytical tasks that it would make errors 'intuiting' the answer to: (have the LLM) write a program and use the output of the program. Note: There is a growing literature in this area, e.g.:

arXiv:2305.04091v1 Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models Lei Wang<sup>1</sup> Wanyu Xu<sup>2</sup> Yihuai Lan<sup>3</sup> Zhiqiang Hu<sup>3</sup> Yunshi Lan<sup>4</sup> Roy Ka-Wei Lee<sup>3</sup> Ee-Peng Lim<sup>1</sup>

#### **4.3.22 'Kasparov Event Horizon' for Object Perception & Handling**

The basic idea of a Kasparov Horizon or Event Horizon is the 'distance' beyond which the AI is blind and fumbles. It is a generalization of how Gary Kasparov eloquently describes the 'distance' or 'depth' beyond which the AI cannot see or handle objects, which is a rather concrete concept if you have interacted a lot with AI either trying to stay within, or trying to stay outside of, this horizon of reach.

Practical applications of this may include the internal project-object handling of LLM models, especially for project-participation and perhaps high-stakes areas such as medical diagnosis, where (e.g. for NLP) if you overload the task with too many parameters or levels the ability of the AI to handle the objects falters.

#### **4.3.23 Scientific Method and Data Science Models**

<https://medium.com/@GeoffreyGordonAshbrook/scientific-method-and-data-science-models-8aa7d6692498>

2023.05.14

The Problem's Main Concepts:

1. How are DS/ML/AI models equivalent to, or not equivalent to, conservative falsifiable hypothesis testing in "traditional" the scientific method?

2. How might DS/ML/AI models represent a forefront of STEM, extending science and stem into new integrated-STEM areas?

3. In the interplay between:

- Probability & Statistics
- DS/AI/ML
- Falsifiability & Hypothesis Testing
- Nonlinearity

3.1 How do all of these fit together in this context?

3.2 Where exactly is: 'the scientific method'?

4. Has the generalized-STEM question been resolved within a timeline of western science and the timeline of the development and refinement of the hypothesis testing, falsifiability, and "the scientific method"?

Overview:

1. People would like this topic to be:

- clean & clear
  - finite and small
  - resolved and confirmable

2. But, and I am optimistic about this, this topic appears to be:

- fragmented
  - large, broad, and with open-ended scope and potential
  - unresolved and with an unknown overall shape

in such a way that there is significant room for expansion into new STEM-integrated, structure-function-ratcheting tools and methods.

Speculation:

DS/AI/ML models are surprisingly equivalent in many cases to the preferred linear model tests for conservative hypothesis testing, which may catalyze a significant expansion in applications of enlightenment-STEM tools, and represent perhaps a new era and conception of how the enlightenment valuation and utilization of ~reason (roughly as expounded by advocates such as Steven Pinker in works like "Enlightenment Now").

#### 4.3.24 AI Bodies and Brains

<https://medium.com/@GeoffreyGordonAshbrook/ai-bodies-brains-solving-a-problem-4f0c59a2c3ca>

#### AI Bodies & Brains: Solving A Problem

2023 June 3, July 14, Aug 2nd, Aug 13 G.G.Ashbrook

AI usually has no body. And, if more peculiarly, AI in a sense has no brain either. This may sound odd as it is probably not uncommon to think of AI as a literal, or proverbial, 'brain in a jar' or at least a brain isolated from a body.



If we are going to do projects involving AI-bodies-and-brains we will need some tractable approach for either dealing with the interconnected nature of bodies and brains, or some kind of clear framework for moving ahead in some direction despite that fact our sort-of-analogy gestures to body-ness and brain-ness tend to fall away into the mists of areas of the world that H.sapiens-humans either do not understand or have largely fantastically imagined to entertain themselves: We do not know what 'bodies' and 'brains' are, or what exactly we are invoking when using the terms. But perhaps 'close enough' will suffice for now. Fortunately the point to be made in this paper does not assume or rely on there being some absolute biology-STEM definition for 'body' and 'brain' (or 'life' or 'intelligence' or anything else).

What we will be talking about here is probably related to ~bodies and ~brains but in a more general sense that cuts across known organism body-plans, colonies of cells or organisms, possible ET forms of life, AI, and novel hybrids involving chimeras across categories, etc., and the unfamiliarity of all that may take a moment to digest. But we will focus on a functional and practical goal relating 'more or less' to something body-ish, with the practical goal being very important and the semantics being disposable when they cease to be helpful.

### **Brain!**

Let's start with 'brain.' In a way, the generative intelligence of gpt-LLM deep learning model artificial neural networks is a containerized sort of hologram of millenia of H.sapien-human intelligence (by whatever name) gleaned from printed literature. (see: Dumbledoor's Portrait paper)

Though we may never understand 'intelligence,' [or "intelligence" may turn out to be a malapropism of meanings as was "phlogiston" (something which we now know does not exist)] a great approach is S.Raniwala's hypothesis that 'intelligence' (as exhibited by GPT-LLM and by H.Sapiens-humans) exists not so much in a homunculus hiding somewhere inside a brain, but in the Firth-relationships **between** concepts in externalized literature: that something of, or all of, 'human-intelligence' exists 'in' language (being a set of connection between parts of language) itself flowing, transmitting, between (or in a network of) a substrate of "people-body-brains." (a Firth-humuculus?) This may seem to be splitting hairs or getting stuck on arcane distinctions, but as you will see below there are very important function, system, design, details for what we mean operationally by 'intelligence.' What exactly can and can't this

'intelligence' do? Or in this case, what is the general relationship between 'intelligence' and a 'body-brain'?

As we will see, we may need to unpack what we mean by 'intelligence' in isolation vs. 'a system that can employ intelligence to guide operations.'

## **Strengths & Weaknesses**

We are in a sense still back in something like the same pre-2020 time when AI had no significant object-handling ability. (See full paper: [https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml) )

*"AI is good at identifying things, but not very good at doing anything with that information."*

While this phrase originally came from describing the difference between a very narrow not-at-all-general AI and the kind of tasks we want AI to do, this still applies surprisingly well when we push ahead with seeing what even object-handling (see paper for explanation of term) gpt-LLM-AI can do in a context involving the following:

- projects
- participants
- interconnecting areas
- generalized STEM

etc. including self-checking, explaining what you mean, following process standards and protocols, and thorough reporting.

(See full paper:  
[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml) )

## **Back to square one?**

The fact that an often powerful 'intelligence without body or brain' can nevertheless often fail to manage math, analytical, system 2 [https://en.wikipedia.org/wiki/Thinking,\\_Fast\\_and\\_Slow](https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow) tasks may be related to our body-brain problem at hand. Perhaps like the genie scenario where you ask for something 'intelligent' and then realize what you got was not what you thought you were asking for, to rehash the immortal story again: back in the 1970's Douglas Hofstadter (in GEB) prepared us for the likelihood that even powerful human-level AI would (and perhaps should) have difficulty with doing math, analytical, system 2 type processes quickly (note:

while I am not aware of Hofstadter was working with Kahneman and Tversky, they both emphasize a similar problem space and in particular fast vs. slow (it was literally the title of Kahneman's summary book). Kahneman and Tversky have helped us to better understand how H.sapiens-human brain-bodies work and quite often fail. And in general most highly intelligent STEM professionals are oblivious to proper project management (perhaps an echo of the pre C.P.Snow pre Sir Eric Ashby years when serious people could seriously assert that a scientist, an engineer, and a mathematician, would never have any business speaking with one another [for a brilliant portrait of this utter tragedy, read Andrew Hodges' "Alan Turing : The Enigma" 1983, for the backstory of how post-war US and UK fumbled for years to build the first generation of computers because they could not overcome this exact social barrier and "leap of understanding"]. Indeed the problem was much worse, with 'science' having no single conception and the various factions comically (in retrospect) fanatical about not recognizing the others (see Ashby, back in print thank goodness). Note, even core events later in computer science and AI are enmeshed with this problem, where none other than John McCarthy (creator or co-creator of the term and current form of AI itself) tried to stop the implementation of none other than the LISP language (his own invention!) from being implemented in reality, not because it was dangerous, but because he saw no conceivable connection between philosophy-thought space (STEM) and reality-engineer-maker-space (also STEM).

[https://en.wikipedia.org/wiki/Lisp\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/Lisp_(programming_language))

History is full of this kind of paradigmatic disparity that makes it so difficult for a given person (or group) at a given time to intellectually-empathize with how the world looks in both the past and future over gulfs of time.

### **Problem & Solutions**

Solving a problem, or doing a project, such as a child's math-word-problem, is not accomplished by a fantasy of 'symbol manipulation' or an invocation of some woolly essence like 'intelligent-ness' alone, however much our please-sensors light up at hearing and retelling such 'just so' mythic tales.

Our challenge is to help the kabbalah-shard-fragment of an Indra's-web of notions that is the GPT-LLM containers of human intelligence to be able to do 'projects' like  $1+1=2$  reliably, by building a sort of 'brain-body' for that AI: an architecture, an OS. a programming language: a debugger, a compiler, a linker. Pick whatever mascots or analogies you like, but we need some framework that can get the job done.

Across disciplines (with a nice recursion around the teaching of language and STEM using language and STEM to teach AI to help teacher's of language and STEM using Language and STEM...so that children can learn Language and STEM...to help etc.) from basic teaching best-practice methodology to Francois Chollet's (he is the creator of Keras the primary AI framework) exhortations of breaking things down into modular recombinant parts that can be assembled and used and re-assembled and re-used, to how many computer languages from C to Rust work, there is a value in breaking down a project-task into parts, solving those, then linking the whole solution back together. And, again see the full paper, when this is a project that involves multiple participants and other project aspects there are many more important parts to this breakdown-build-up pattern that do not seem readily apparent when it is considered a mere flourish to adorn a single person solving a hypothetical problem in their mind (or to imagine and believe they have solved it...).

Probably not for the last time, the system by which we might furnish and AI's echos of literary intelligence with a 'brain' are strikingly similar to how we teach...or try to teach...children how to think, write, and solve problems in an organized effective way (with 'externalization' being a stubborn theme). And interestingly even just the disembodied AI-echo that is GPT-LLM shares many similarities with the strength and weaknesses of corporeal H.sapiens-humans. See: <https://medium.com/@GeoffreyGordonAshbrook/ai-counting-problems-8cb9f66e4c7f>

And there may be more to this than merely saying the former is an echo of the mind of the latter: there may be a more general set of process dynamics going on here, beyond contagion of idiosyncrasies.

Does this sound familiar?

- *Use a structured framework.*
- *Make a plan for what you are doing.*
- *Understand the focus and scope of your task.*
- *Follow your plan.*
- *Show your work!*
- *Do not skip steps!*
- *Write everything out clearly!*
- *Explain what you mean.*
- *Double check your answer!*
- *Proofread your work!*
- *Revise your work!*
- *Publish, share, or submit your work.*

For years this has been the nagging guidance of H.sapiens-human adults to H.sapiens-human children, and now it is the nagging guidance of H.sapiens-humans (of all ages) to an AI (which is a fragment of their own minds from ancestral texts).

But how is AI, having no body and no brain, going to do this?

We will build something with which the AI can do this, and this something is, or may be akin to, a brain in which this AI, which is so good at making observations, can then actually do something with that information.

### **Intuition vs. Code ... "Show your work..."**

Another important strength-weakness node in this approach is the observation that while gpt-LLM models are often very slapdash about estimating an overall answer, they can be extremely good at writing computer code to do the same exact thing properly. On the one hand this may seem strange, or seem like a deficiency, but this is remarkably similar to how H.sapiens-humans behave and it may be part of systems in general, (may be) for example if the 'system 1'(non-analytical or intuitive-guesstimating) and 'system 2'(analytical) of Kahnman and Tversky are a more general part of the world and not just an accident of evolutionary history for one branch of mysteriously hairless ape.

Another possible context for this distinction (and another excuse to possibly merge chess mysteries into an AI narrative) is strategy vs. tactics. Or is that a different subject altogether? All in all we are still rather new to this whole space of general-STEM, game-theory, problem solving, nature-of-language, nature-of-thought, so hopefully a few hundred years from now people will have a better idea of what's going on and they can safely look back at our confusion with smug disdain. Bless them.

## **Part 2 : Solving A Problem**

As a concrete example, let's look at the process of solving a multi-step problem, or solving a problem with an externalized project-share-able and scrutinize-able, error-check-able, process. This description could still refer to an unhelpfully large number of problems and kinds of problems (or a very large problem-space), so

let's start with some concrete first samples (however they are pigeonholed):

A math problem:

A math word-problem: A conventional k-12 school question.

A logical-puzzle problem: Who is coming to the party?

Note: (See a later paper on modularizing problem space with an example 'counting problem' in this same overall context.)

## Part 3 : Math, Language, STEM, Computer Science

### Riddle Me This

quote by Francois Chollet:

Page 26, Chapter 2, opening paragraph,

<https://www.manning.com/books/deep-learning-with-python-second-edition>

<https://www.amazon.com/Learning-Python-Second-Fran%C3%A7ois-Chollet/dp/1617296864>

*"The most precise, unambiguous description of a mathematical operation is its executable code."*

This perhaps simple seeming statement is somewhere between nuanced, subtle, bold, and profound, as both the long history of debates over whether mathematics and logic has any relationship to STEM and practical problem solving is tumultuous and largely unresolved, and computer science (discrete or nonlinear) is rarely mentioned as part of STEM at all: so while the time may have come to postulate an authoritative seat for computer science to best define a very real ability for math-logic to solve practical STEM problems...most of the rest of the world is still bickering over preclusive component objections.

<https://www.amazon.com/Technology-Academics-Universities-Scientific-Revolution/dp/1015312659/>

A truly fascinating lineage of books in and about the history of thinking about STEM and science is the progression from Sir Eric Ashby's "Technology and the Academics" to C.P. Snow's "The Two Cultures," to Stephen Pinker's "Enlightenment Now." Though each of which pays homage to the previous and states their goal to re-write, update, and build upon the past, there is a 'arc of history' of sorts from the first two books that trace out a bewildering diversity of disagreements over whether there is or could be a generalized or unified STEM and how that relates to the world, with the perhaps a-touch-too-combatively glib Pinker's work making his cut-to-the-chase case that everything is reason and anyone who disagrees is a dangerous moron.

### *"someone else's mind"*

Along these lines, unassuming books such as "A Programmer's Introduction to Mathematics: Second Edition" by Jeremy Kun (Author) <https://www.amazon.com/dp/B088N68LTJ>

may contain important practical insights for how 'tricks of the trade' for both math-logic and programming can come together for systematic problem solving.

From chapter 2, page 12:

*"This theme, that most mathematics is designed for human-to-human communication, will follow us throughout this book. Mathematical discourse is about getting a vivid and rigorous idea from your mind into someone else's mind."*

Whether or not these aspirational statements by Chollet and Kun are true-enough remains to be seen, but the game is afoot as we set out to transfer ideas not only between 'human' minds (whatever that means), but between any mind or intelligence in general, and even using decidedly non-mind tools: (perhaps similar to the Raniwala Hypothesis that conceptual-intelligence exists in language space (in the measurable Firth-space between n-grams) not in homunculi hiding in brains,) 'mathematics' (which Skip Ellis believed was a natural language (another debate that I cannot resolve)) is contextualized as being able to contain and convey ideas (and/or that computer-code is contextualized as being able to convey mathematical ideas between humans and machines) and in language-space. (To shorten that: The claim that 'mathematics' is able to contain and convey ideas in language-space.) This is testable and falsifiable, and we shall see what works and what fails.

Examples and Possibilities:

Again, in practicality, there will probably be a slowly or quickly growing set of types of problems that can be systematically broken down and solution-module-compiled in a general STEM etc. approach, with the open-ended "any problem" being too open-ended to resolve.

As with any area in computing, 'programming language' 'compiler' 'operating system' 'network architecture' 'processor architecture' 'gpu' 'tpu' 'memory storage' are categories with perhaps infinite possible solutions and variations (either with pluses and minuses) usually for specific implementations and uses, these not individual specific things of which there is only one example.

#### Part 4 : Context

On the one hand it may sound risky to 'give an AI the ability to do anything!' but that is perhaps an overly vague, hyperbolic and inaccurate way of describing this small step into a larger understanding of the nature of STEM, projects, and our own intelligence in a jar: this has not, per se, solved or even addressed

questions like 'volition' and 'independent action' and 'appetition (or desire)' etc.

This proposed architecture is an additional layer to the same conventional 'chat-bot' interface so that some narrowly-defined types of questions asked of the chat-bot can involve an incrementally-less superficial ability to interface with problems, and projects, and object-relationships. But, back to the title, there still is no mature body or brain. And, if frustratingly, this is not a 'do everything generally' theory-of-everything.

It might be interesting to experiment and say to such a semi-embodied-AI, should it be possible to build and craft one using such a mini-brain architecture: "Make a list of things you would like to do today, do the item at the top of your list, and report back! How did it go?" And this may pre-suppose that the AI can 'do' things in 'the world.' The version of gpt4 I have now still says it cannot access the internet in any way and its knowledge is limited data from two years ago. Politics aside, adding in more abilities for the ai-fragment use, for example, to look up information on wikipedia, will eventually be created (and permitted).

But even this example, which seems like the AI "deciding to do something on it's own volition" would still be a single process, run in isolation, that would terminate at the end of the instruction, and the AI would return to its amnesiac state, passively inert until something bumps into it that it reacts to; still an image of intelligence (however high resolution) that can only reflect and refract back from a prompt shown into it: There is no body.

The Mirror Analogy: "Mirror Mirror On the Wall..."

We might look at GPT-LLM AI as a kind of intelligent mirror; it can reflect intelligence; it can reflect articulately. But that one, passive, reactive, involuntary, reaction is all this fragment can do. For example, and hopefully this will be done some day,

<https://medium.com/@GeoffreyGordonAshbrook/our-ai-ancestors-dumbledores-portrait-and-ray-kurzweil-s-father-85ec89f85224>

you can imagine an intelligent mirror that has been trained with all of the works and words (and whatever other data) of your wise grandmother, or some benevolent great thinker and teacher from the past like Niels Bohr. You could hang that mirror on your wall, and give it things to 'reflect on.' And, if well-made, it could reflect with great wisdom, subtlety, nuance, insight, depth of experience,



and articulation. But it would not be a person, an organism, an ecosystem, it would be a fragmentary reflection no matter how profoundly articulate.

For example a mirror that can only reflect cannot solve a multi-step problem. You may visualize this as a contraption with laser light sources and mirrors and lenses. This ai-mirror may be able to give (reflect) an all-in-one guesstimation as to what the outcome might be, but that would be (as Hofstadter predicted) a lagging ability (a fast-fuzzy-guess at what the outcome of a longer slower process probably might be), or an an internal guesstimation ability not to be confused with the outcome of an externalized process. The mirror can reflect on each step, but it cannot alone do anything with the steps for a problem. It cannot feed a step to itself. It does not know what step (in or out of sequence) it will reflect on next. It can be fed a jumble of randomly ordered instructions taken from unrelated projects and forever patiently reflect brilliantly on each one in term, with no memory or concern about what came before or what comes next (perhaps a Hume-moment-sequence nightmare...Hume's Daemon if he does not have one named after him yet). In some ways this is like a CPU or ALU in a Turing/Von-neuman type computer architecture, processing (or reflecting on) one operation at a time, passively (and not fussed about whether the concurrency and parallelism and swapping of instruction sets is being managed properly). It cannot refuse to react. It cannot 'react' unless there is something to react to (or reflect upon). A CPU is not a computer. An intelligent mirror is not a person. A brilliant gpt-ai fragment is not a brain.

A computer, on the other hand, with an architecture and operating system of some kind, and many peripheral components, is more of an 'organism' that can knit and link steps together. A CPU or ALU cannot compile a long program into a finished executable binary in a single boolean operation (with a running theme here so far unmentioned of how much information can or should be squeezed into a single operation...a single reflection, a single 'prompt'), but a full computer body-brain can link and compile the parts of a program. A mere reflection of a person could not participate in a project, with bits and parts of schedules and instructions being crafted and passed back and forth, multi-step tasks accomplished, plans followed and reported on, plans changing; but a "person," or participant, with a body-brain can do these things. (Or can, in an ecosystem of other participants?)

Using this analogy to frame our problem-problem, how can we build a 'brain' around this intelligent mirror so that the intelligence that

is only passively contained can be channeled and embodied into an architecture that can, if only one small step, solve a problem. The wise-mirror can do each step, and knows how to link the steps, but it needs a body to do so.

There is also perhaps an interesting twist or fold, where on the one hand we think of AI needing extra furniture to be more like 'the solitary and sovereign human brain,' but H.sapiens-humans are exemplars of the gravity and tendency to not show their work and to jump to opaque emotive reactions, making suspiciously similar mistakes as compared with AI, which on the one hand is not surprising because the AI is literally a reflection of human behavior, put perhaps a bit more surprisingly this is an architecture question not an intelligence question, perhaps suggesting that H.sapiens-humans as well have an intelligence-vs-architectures set of problems in themselves, which in that light may make sense of the fraught relationship between H.sapiens-humans and their own bodies and minds, societies, projects, etc.

#### "The Spitting Image"

Probably most people have often heard people talk about a child in the family as the 'mirror image' of some, often long departed, relative. And indeed, on a large enough display of family photos without dates on the photos, it can be impossible to tell whether a picture is of one person or of their grandmother at the same age. Homo-sapiens humans are in some ways mirrors of past iterations of their amnesiac-phylogenetic-body, and in a way that is curiously complimentary to the AI mirrors of our ancestors. Biological reflections have volition and bodies and minds and interact with the world, but they have (at least not superficially) none of the memories, experiences, and ideas, of their ancestors. The AI-mirror on the other hand has only those things: memories, knowledge, wisdom, perspectives, etc. Is this a match made in...heaven?

#### Part 5: Process

An interesting if oddly invisible part of the history of computer science is the 'invention' of the idea that computation can be done by a process, in particular an externalized process.

And in assuming that a generative AI is 'internally using an externalized process' while also not addressing generalized STEM or

how people solve problems is a fascinating mixed-up half-amnesia where we both assume that problems are solved by analytical steps, forgetting that this was ever not-obvious, but also forget that this distinction exists assuming that generative internal type 1 imagination is type 2 systematic process. What could be a more 'human' assertion?

This may even go back, in a way, at least to the not very clear 'origin' of algebra as a formal process of defining variables, which only very slowly became formalized and adopted (perhaps too slowly for anyone to really notice that it happened) between Al-Khwārizmī and Girolamo Cardano. (see "Significant Figures" by Ian Stewart)  
<https://www.amazon.com/Significant-Figures-Lives-Great-Mathematicians/dp/0465096123>

Also see:

- [https://en.m.wikipedia.org/wiki/Prompt\\_engineering](https://en.m.wikipedia.org/wiki/Prompt_engineering)

**4.3.25** Interpreting Hofstadter's Gap: AI, Music, Math, Language

<https://medium.com/@GeoffreyGordonAshbrook/interpreting-hofstadters-gap-ai-music-math-language-3e17d37a3d5>  
**Interpreting Hofstadter's Gap: AI, Music, Math, Language**

2023.07.21,08.10 gg.ashbrook

## **I. Music, Math, Text**

In a rhyming dance of quandaries through the ages and perhaps exemplified by GEB [and perhaps Finnegan's Wake, ...maybe Tristram Shandy?] is now manifesting in a learning gap that may be an invisible stumbling block for AI. In short, the often literally pictorial figures used to show non-text symbols for math, music, and other areas in most published texts are, being 'non-text', invisible for either discussion with AI such as GPT4 and possibly even as training data (though exact details of training data are unknown). For example, many wonder why the math skills of AI are especially poor; part of this large and interesting story may be that much printed math material falls outside of possible training data (pictures, not 'text').

While it may sound surprising that we have not finished knitting together math and language and computers and all the other STEM areas, for all reasonable intents and purposes, we have, at the same

time, in various ways not yet committed ourselves to recognizing this as a goal.

### "Terrified"

In the fabulous introduction to Melanie Mitchell's great book on machine learning, she recounts a poignant story (that I would not be able to do justice to here, please buy her book!)

<https://www.amazon.com/Artificial-Intelligence-Guide-Thinking-Humans/dp/0374257833/>  
about this set of topics coming to a head and leading to some unresolved confusion when Douglass Hofstadter was invited to speak to researchers at Google. In very short summary, Hofstadter was concerned about directions that might be taken or revealed at the intersection of math, music, and AI, but in a way that confused the researchers who had invited him to come and share his thoughts. He was worried about what we might discover about the nature of, for example, musical patterns in the human world, whereas the researchers were entirely focused on studying the AI-side of things (not worrying about what that might reveal about biology). (The real story being of course much more interesting than my procrustean abridgement...)

In what in other contexts may seem merely an oddity, during medieval music studies there was a time when the study of 'pure music' was so compartmentalized from performance that it became a kind of sub-set of math-logic-computation. (Note: if you have not, take a read through Kepler's original writings on the structure of the solar system and you may be in for a surprise or two.) On the one hand this extreme stance on pure vs. applied music studies may seem a kind of neutral or pejorative decadence to require 'serious' music to exist only in formal logic and math and never to be played (at least from the point of view of someone looking to buy tickets to the orchestra), on the other hand this history seems at odds with our present day pickle of not being able to combine music-math and math-math, with computer-math. What kind of a gap could there be here?

(Prepare yourself for another extreme summarization...) We train AI models such as LLM-GPT (at least in 2023) on texts and the AI makes a thought-intelligence-concept-map by looking at the relationships between all the terms, ideas, concepts, etc., in those texts. But this only works (as of 2023) for text...not pictures or sounds, or hand gestures, or anything not printed in characters in linear sequences on a document. And while music and math have evolved symbol-systems, those are very-usually not expressed in documents as linear sets of language-characters, not 'serialized,' which means that AI cannot see them any more than they can see the facsimile

plate of a Cezanne bowl of fruit on page 100 of an art-history book, though it digests any description and discussion there may be of what is on that plate.

The history of the evolution of math notation, and the movement to make expressions of math more standardized and formal and rigorous, is long and is a story we are still in the middle of. Both Alan Turing and John Von Neuman (two figures at the head of conceiving of, and designing, and building, and programming, the first generation of digital computers) began as mathematicians assigned to work on

"[Hilbert's Problems](https://en.wikipedia.org/wiki/Hilbert%27s_problems)"

[https://en.wikipedia.org/wiki/Hilbert%27s\\_problems](https://en.wikipedia.org/wiki/Hilbert%27s_problems) which, like Russel and Whitehead's project (intersecting with both Godel (and Church) and Turing) to, if I am permitted to contextualize this way, move toward a generalized or unified STEM where math, logic, the sciences, engineering, technology, medicine, etc. (and later computer science) would all be fundamentally compatible (speaking the same language as it were), though at the time (around 1900) the stated goals of Hilbert's challenges were probably phrased and paraphrased more modestly as seeking some technical overlap between logic and math, and tidying up loose ends not formally defined, as a grand generalized STEM was (see Sir Eric Ashby

<https://www.amazon.com/Technology-Academics-Universities-Scientific-R evolution/dp/1015312659/> & C.P. Snow et al) too mind-blowing to publicly admit to thinking about. (In the 1950's both in the UK and US this inability to conceive of connections between STEM areas was a significant obstacle to actually building the first computers which involved people from many STEM areas actually talking to each other and working on the same project, which proved beyond human comprehension in principle at the time (see Andrew Hodges' "Alan Turing: The Enigma")

<https://www.amazon.com/Alan-Turing-Enigma-Andrew-Hodges/dp/0671492071/>

Part of what makes GEB such a rich spring to return to is the bold way in which Hofstadter weaves these themes and music and math and the richness of language together in questions fruitfully, as a wellspring of AI and of humanity, not posing the intersections as a stumbling block precluding any future progress.

Is it not strange that the very topic, the very intersection, that GEB focused on to electrify minds all around the world and resurrect the AI movement after multiple proverbial planet-sterilizing impacts of funding cuts and...boating accidents, mysteriously poisoned

apples, and all manner of misfortunes, the very intersection that birthed the 'music of the spheres' science of astronomy, the very intersection that set the 1900's on its Hilbert course with destiny, and also the same intersection that 'terrified' Hofstadter (and utterly baffled Google), and is now the same intersection that confounds the cutting edge of AI intelligence...is this all not very oddly all continually orbiting around and echoing off the same intersection-gap?

## **II. Hofstadter's Gap**

Is there a 'gap' of some sort here, or is this a simple translation process that happens to not have been solved and automated yet? Is there anything more here than just doing the work to create processes to connect these dots, basically making a text-serialized system to contain the less linear picture-diagrams of "symbols"?

### **Representation in a context of externalized representation:**

In music performance one of the main factors is how many people are performing together and how are they coordinating, and depending on the music tradition there are various "systems" of procedures and rules by which different types of instruments (including the human voice and percussion which may also be just the human body) are able to coordinate, often (though this has largely be forgotten in western classical music) with a balance between improvisation and a shared framework.

### **What does the representation contain?**

In the case of computer code, the representation must be complete (or call specifically for other functions that are needed). In the case of music improvisation the code is usually deliberately incomplete: sufficient for use but not the entirety of the process.

Math, interestingly (likely a controversial subject I am naively wondering into here, sorry about about) appears historically to be somewhere in between. The reputation of math in the west is often the super masculine hyper-serious high-brow noble strict posturing, with affectations of being as complete as assembly language. But historically and across various traditions (e.g. traditional japanese geometry puzzle-exchange as part of some shito-shrine traditions, or renaissance Europe's public testosterone math competition battles judged as sports to a crowd)) much of math seems to be more like music, more a recreational social dance than a maker-STEM engineering-project. E.g. for most of history integrating math with

STEM was something many mathematicians fought and despised, taking pride in their purity in separation from worldly parts of the world, even up through WWII with a leading german mathematician taking pride and reassurance that is life's work was of no use to anyone (specifically meaning not useful to the military in this case).

So we still have questions about Hilbert's questions. Hilbert, I conjecture, though I may be wrong, wanted (or would have wanted) math to take its place as part of STEM, to be testable and clear and not only an arcane branch of pure poetry where leaders in the field did not, and were not, expected to explain and show what they claimed and did. But is this STEM-Math all of math or just part of math, all of the math-mind-imagination or just a useful edge of a larger less confined space?

If there is (as surely there is) improvisational music, and if music is hip-together close with math, is there not in some way improvisational math (can the two be completely separated)? Yet does it make sense for improvisational math to be required to be in the same category as STEM machine-code type math?

As with music, you can create a rigid, fully-defined, system of exact and unyielding frequencies and volumes and tempos and official terms and techniques, but the more strict that system then the fewer actual world music systems, arts, skills, and traditions, will be compatible with such a 'complete-framework.' And there are a lot of formal traditional music traditions out there.

I am very much not arguing, as some people of a more post-modern, or social-constructivist, or this or that school of thought, may argue, that STEM is somehow 'just' another soft social free-for-all merely pretending to have engineering precision. I am asking more about the overall space of definitions and projects.

While there does not seem to be, or need to be, a one size fits all mega-STEM framework that defines everything in existence, that is not the goal or meaning of a generalized-STEM. Connecting the various areas of STEM is just this, connecting and translating between expressions in each area of STEM. Being able to express and translate between and across expressions in math, logic, music, philosophy, physics, engineering, medicine, statistics, chemistry, biology, computer science, does not pre-determine what you are doing with those expressions (or even if you have any idea what you are doing).

Perhaps going back to a music-analogy, sometimes improvisation is the goal.

Sometimes we are not trying to define a method (to find an answer to a question). Sometimes we are not trying to define an answer (to a question). Sometimes we are not dealing with a question or problem at all.

There is a strong attraction to familiar analogies. In the 1800's it was clocks and steam engines. In the 1900's through 2020 it was digital computers.

While it is important for us to understand assembly language running on a cpu, the world is a large place full of ecosystems and signals and generation and equilibria and many forms of system dynamics that we are still struggling to begin to map. And while writing a bare metal system to run on architectures may knit together known STEM, it does not bring us closer per se to the larger world we need to understand, and in which we are living perhaps or perhaps not planning to survive in. And this old teeming canyon of Hofstadter's Gap may be a compass heading for what our eyes can not yet focus on.

### III. Frontier

We may be tempted by circumstance to orient ourselves toward Hofstadter's Gap in such a way that this feature of the world is something unsightly to be resolved and fixed and solved and hushed up and put behind us and smoothed over, but is this gap a potential?

Could Hofstadter's Gap be not a human error in procrastination in not yet finding tidy closure, but more in the spirit of GEB, an endless frontier of convoluted quasi-intersections? Is this rather a potential that fuels not only wonder but the mysteries of creativity and the progress through life's largely imponderable quandaries, through all the unknown fissures of T.S.Elliot's 'heap of broken images'?

Is this not a mistake or obstacle but some kind of device in and of itself that we do not yet understand? Or some rich riparian zone like the "Canals" of Mars that were once imaginatively if naively thought to be waterways and the site of life...and then later in a very different and altogether more proper and scientific way...thought to be waterways and the site of life! Is this a kind of door into the



hollow-earth of the fabric of Charles Dodgson curiosity-patterns...of how STEM, like four valent carbon, is somehow so creative a set of linkages?

For example James Gleik opens his "The Information" with a discussion of traditional African talking-drums in a context of internet-like signal exchange protocols and information theory. The intersection of music and signals and networks and encryption and decryption, is not meant to be a finite 'solved' system, but a working and open ended platform. Above we discussed how the even tempered scale is not a signal protocol that works for many traditional forms of music. So what then is the science of working protocols? The world is full of working platforms, platforms that allow for improvisation and experimentation.

#### IV. Generalized STEM, Generalized Projects, Generalized Collapse

There are both more intersecting connections and more incongruous gaps than we often suppose at a given time. While in 2023 many more people than in 1923 are likely open to the idea that a mathematician, an engineer, and a scientist may appropriately be permitted to speak with one another (and beyond permission that they obviously should and must in order to get anything done), there are still many vague, empty, and undiscussed areas of a larger problem-project space of unified, generalized, STEM that remain unpopular.

Those focused on in the report that this article is hoped to supplement include:

- Projects
- System Collapse
- Ethics
- Participation
- Learning

And there are no doubt other areas that will in retrospect seem obvious yet at present do not come to mind. For example, people looking back on year 2023 hundreds of years in the future may find it inscrutable and inexcusable that civil and criminal law, and education, are not considered compatible with logic and testability. Or on the other hand, perhaps they will have evidence to the contrary.

## V: Equations as Numbers

Even the question of 'symbol manipulation' in the history of computer science, which may go back to the laudable and foundational writings and foresight of Ada Lovelace, should lead one to...have a pause. Still rather awkwardly in the 2020's we rotate the cannon and dutifully recite: "computers manipulate symbols and anything that can be represented by a symbol can be manipulated by a computer, in heaven's name we pray," despite that this forces us into ridiculous twists of language and logic, for example requiring 'symbols' to refer only to relative electrical currents as "physical symbols." ...What?

But while it is ridiculous to perpetuate the narrative that all of the worlds' mostly magical symbols can simply be fed into a computer and computed with, the perhaps similar-ish topic of handling functions and equations as numbers (not as sigils), is more to the point.

What does it mean to have a discrete and unit-test-able function that returns not a value but, in a sense, another function, or an equation?

In a way this brings us, as usual, back to the 1930's when Kurt Godel, Alonzo Church, and Alan Turing were all working on Hilbert challenges and consequently discovering and inventing aspects of computer science. e.g. Kurt Godel's "Godel-numbers" are numbers that refer to functions (roughly). Alonzo Church created what is now known as 'functional programming' (and the "Lambda Calculus" of functions) where functions are the unit of computation and one can speak of a function that inputs and outputs other functions. And Alan Turing created the paradigm (often attributed to John Von Neuman (who helped Turing to get his start in academia though the communications (if any) between the two are mysteriously few and lost to WWII secrecy)) of having computers mix code-as-numbers with data-as-numbers together in the same "Von Neuman" computer architecture (roughly stated).

## VI: Categories of Types of Signals and Responses

As a general place to start we might assume that Categories of Types of Signals and Responses follow categories of type of systems, from system and definition behavior studies; roughly:

- logic-math
- statistical/probabilistic interface

- one-tree of core linear descriptions of physical phenomena using statistically-bridged math-logic
  - dynamical systems connected to core linear systems
- Perhaps with the turing-church twist, that logic-math can be either equation/function or value.

And while it is rare, we may need to add quantum and relativistic edge cases if they are too irregular.

There is the core CS history-mystery of Hilbert-Completeness and Intermingled instructions with data:

- Kurt Godel (often not mentioned as a direct contributor to computer science)
- Alanzo Church
- Alan Turing
- John Von Neuman

#### **Instruction vs. Data** (or not a dichotomy?)

The intermingling and interaction (sometimes unintended) of low-level-instructions for the computer and higher level data to be handled by programs is most usually characterized as a dichotomy. But perhaps this should be broken into more categories that cohabit in information substrates?

1. machine/operation instructions
2. functions/equations
3. variable values (strings, numbers, arrays, etc)
4. bridge-translation-representations (e,g. Godel numbers?)  
(a type of machine instruction? ...not an operation...? )

## **VII: Training and Learning**

While there is a fair quantity of commentary about the inability of AI to answer math questions, there is suspiciously little attention paid to at least two points.

1. There is no way for LLM-GPT AI to see the math literature to train on.
2. Even though there are probably millions or billions or maybe more than that datasets available or at least mentioned online, as of 2023.07 the entire topic of a math question answer dataset is utterly nonexistent.

The entire process of interfacing AI with math has not even begun...in late 2023...and the often very credentialed people who are critiquing AI-math ability have somehow not noticed this, carrying on as if there is some phantom math-ai industry.

So much commentary about why "other people" have not somehow trained AI to do math, and also a complete blank on even the most basic initial steps: A ponderous gap that keeps on gapping.

### **VIII: Neighbors and Connections**

Three or more areas that bare a striking resemblance yet which are nevertheless separated by gaps are:

1. cellular automata & neighbors (with discrete rules)
2. matrices (without neighbors)
3. subsymbolic artificial neural networks (entirely made of neighbors, but with importantly nuanced rules (neural networks in 2020 vs. 1950)
4. digital memory storage (usually without neighbors)
5. 'Tapes' of instructions (as in a two dimensional linear turing-automata)

In some ways these systems are so similar and parallel and analogous that they appear to be obviously part of one archetype or mode of system, but at the same time a single implementation that combines all the properties of all of the above is nonsensical. Here again we have a working platform with fruitful gaps, and recurring themes.

See:

<https://www.amazon.com/Artificial-Intelligence-Guide-Thinking-Humans/dp/0374257833/>  
[https://en.wikipedia.org/wiki/Hilbert%27s\\_problems](https://en.wikipedia.org/wiki/Hilbert%27s_problems)  
<https://www.amazon.com/Alan-Turing-Enigma-Andrew-Hodges/dp/0671492071/>  
[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)  
<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/1324003995/>  
<https://www.amazon.com/Fancy-Bear-Goes-Phishing-Extraordinary/dp/0374601178/>

For more discussion of theory and practice of the mixture of machine instructions and information values, or instructions vs. data, see:

1. "Fancy Bear Goes Phishing", by Scott J. Shapiro  
<https://www.amazon.com/Fancy-Bear-Goes-Phishing-Extraordinary/dp/0374601178/>

2. "The Man from the Future, The Visionary Life of John von Neumann", by Ananyo Bhattacharya

<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/1324003995/>

#### 4.3.26 Modularizing Problem Space for AI

<https://medium.com/@GeoffreyGordonAshbrook/modularizing-problem-space-for-ai-following-a-wedge-by-sight-ab88796c4b57>

Modularizing Problem Space for AI:

Following a Wedge by Sight

Let's follow a wedge of thought, by line of sight, into future near and far.

Three Steps:

1. Adding and Subtracting
2. Who is coming to the party?
3. The Nth Sub-Supplier On The Nth Contract

## Modularity & Performance

1. Adding and Subtracting

see: [https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml/blob/main/research\\_testing\\_labs/Modular\\_Counting\\_Problems\\_v5.ipynb](https://github.com/lineality/object_relationship_spaces_ai_ml/blob/main/research_testing_labs/Modular_Counting_Problems_v5.ipynb)

Can we construct a modular problem (or a way of modularizing problems), where changing one (modular) aspect of the problem (without changing the solution process at all) somehow interferes with the solvability of that problem? [This paper argues: Yes.]

With this example ([see the notebook link](#)) it will hopefully be visible how a problem becomes empirically less solvable by a bare passive-reflective AI (for discussion see other articles, 1, 2 ,3, etc.), even though the solution process has not changed at all, and where we are not talking about extreme cases that stress the limits of computer hardware memory storage. Rather, for example, a question with module-size 25 is answerable, but module-size 50 is not.

*This is a (hopefully) small and intelligible step into a class (a set) of problems where we can see how changes can be made to a problem that, without changing anything about the problem in terms of the steps used to solve it, the...*

*This is not a question of computational "complexity" or limited resources. This is a much more understandably-modest change from dealing with, say, 5 things to dealing with 10 things, or from dealing with 10 things to dealing with 20 things, in exactly the same way following exactly the same steps; yet something is breaking down. Something is falling outside the range of perception.*

This idea that we can create problems that are modular and adjustable so that we can measurably change aspects of the problem that affect empirical-commutability is hopefully useful as we strive to deploy in real life empirical computations (that, hopefully, as in our self driving cars or office-coffee-bots, or meeting scheduler-bots, are able to accomplish their tasks without (much) calamity). (Or can we eliminate calamity without eliminating hilarity?)

In this notebook, which you can open in a colab:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml/blob/main/research\\_testing\\_labs/Modular\\_Counting\\_Problems\\_v5.ipynb](https://github.com/lineality/object_relationship_spaces_ai_ml/blob/main/research_testing_labs/Modular_Counting_Problems_v5.ipynb)

you can interact with the fully modular sample problem, where the 'base-problem' or 'root problem' (or whatever the clearest description ends up being) does not change, and the solution-process does not change, but the problem relies on an input-module that has different measurable characteristics: in this case an arbitrary list of colors, but it could be any words or indeed any list of string-elements (perhaps unicode characters or any-code characters). The main property of the list is how long it is, but other factors may be worth investigating as well e.g. how long each of the string-elements are.

Let's say there are only two strings in the list, but each is a Charlotte Bronte novel...would that confuse ai? Again, the steps to solve the problem are identical (and trivial, just a few basic functions: get the first letters, count, add, subtract, done!) the problem is completely deterministic (unless you put quantum states into the string behind a black hole event horizon or something...hawking radiation stings? Usually this will be completely deterministic.)

edge cases:

- what if the lists are empty?
  - where if there is only one item?
  - what if all the items are the same?
  - what if all the items are variations on the same thing?
- (Hofstadter collected cases of the letter A, we could make strings different versions of the same letter...)

- somewhat changing the problem, but adding more lists, or lists of lists (just a slight tweak to the solution-process)
- peculiar repetitive patterns in those elements

Some edge cases may be outside of a 'normal' problem (such as empty lists which arguably break the solution functions depending on hardware-software details. However, in addition to pipeline problems that are not AI problems but issues with processing what goes into or comes out of the AI, once those are dealt with there may be AI issues as well.

There eventually will be character-set issues which become monstrous as you drift away from ascii-land) but other 'edge' cases such as having a "large" number of results (not numbers that are so large they break the computer memory storage systems) should be perfectly valid.

*Reading Question 1: Are counting and adding the same thing?*

## **Participants & Roles:**

2. Who is coming to the party?

A more familiar but comparable space.

Hopefully the modularity and automate-ability of the above problem can be seen, however the problem itself is not very user-friendly.

For example, after spending just a few minutes away from it, when going back to double check that the wording of the question contained no typos I felt like I had to contort my mind in un-graceful-ways to model what is essentially an extremely simple set of problems. (find first letters, count, add, subtract, etc.) And then of course there is the "And why exactly am I spending valuable time on this pointless problem?" question, which is aggravated by the arbitrary-ness of the problem-question.

So let's try to bring these concepts a step closer to empirical situations that language can more easily refer to, say, attendance at a social gathering: Who is coming to the party? This is a kind of simple problem space in the Woody Allan world where the only clear task is bothering to show up at all (and where all else is moot).

Even in a space as simple as one where 'attendance' is the only task-outcome being considered, we can create diverse modular factors

upon which this simple binary deterministic outcome is based (and each with possible variations like the first example where modules can be changed with the solution-calculation being unchanged, but where the empirical solvability does change).

You could even make an analogous version of the party question, by giving a list of possible attendees and giving some ridiculous rule where the only people who can come to the part are Z number of people counting from the list of the party...where z is the solution to the above count-add-subtract problem with the list of attendees as the input. The only additional step is to apply that number to the list, slice the name-list for that number, done.

## **Projects are Problems**

### 3. Integrating The Nth Sub-Supplier On The Nth Contract

But life isn't all about going to birthday parties (...[or is it?](#)). How can we follow the wedge a bit closer to the elusive accouterments of everyday life...

Problems can be a hobby, or sport of riddles, or "symbol manipulations," or whatever fancies and adornments we whimsy, but project-space is also a problem space; and how these two realms of problems empirical and problems philosophical come together in the same STEM riparian zone is one slice of the slowly unfurling story of life on earth and the discoveries of the strange intelligences growing there.

In this third step we will move an increment closer to projects, which brings us into the arenas of project-management and logistics, and also into the General-STEM area where, for example, mathematicians and linguists and engineers and ecologists and accountants all work on the same measurable-module based projects (with outcomes for users and stakeholders).

There is (if not obviously) an incomprehensibly vast spectrum here so let's start with a minimal MVP version: two participants on a project, say, to supply variables to the above-party question.

One of the key modules we can introduce here is the marvelous inch of time: schedules.



Note: Scheduling is a delicious set of real world conundrums (aside from the fact that H.sapiens-humans seem to have evolved to be especially bad at scheduling). And in a sense scheduling problems are a set of "Who is coming to the party?" problems for each possible time for the scheduled event. Who can work the opening shift Thursday? Who can come to a 10am or 2pm meeting, and so at which time should we hold the meeting? When will the parts all be done for a product roll out? Who is coming to the party?

### **Modelar Schedule Project Problems:**

Let's try to start with a miniature problem that is understandable and at least starts to cover a new area. In AI there is the class Woz-AI test of making coffee...so let's do a variation on that involving:

- multiple participants
- a schedule
- tea (sorry)

We might think of this as an intermediary project problem, or a single modular project problem, as it does not include the entire project process from start to finish including the overall outcome and after effects of that outcome, but perhaps such large projects will inherently fall into a hierarchy (or web) of interlocking modular problems and projects that can be looked at separately. In this case, the mini-project outcome is to make sure everyone on the project team gets their tea. Again, we could add as many sub-modules of additional factors as wished, but let's try to start with a most-basic example.

Office Tea-Bot Problem: (a Woz-Test variation...in tea major)

There are various questions that can be asked with a simple starting input scenario:

Input is two dictionaries:

1. a dictionary of when types of tea are scheduled to arrive
2. a dictionary of what type of tea each employee drinks

```
tea_schedule={'Assam': 'Monday','darjeeling': 'Tuesday'}
employee_teas={'Betty': 'Assam', 'Tom': 'darjeeling'}
```

Question 1:

How many people can arrive on X day?

(variation: what is the difference between attendance on X day vs. Y day?)

Question 2:

Who can attend a meeting on X day?

(Variation: Can employees A and B both attend on day X?)

Question 3:

Assuming Employees will not attend a meeting without their tea, what is the first day of week when all employees can attend?

### **What Colour is a Strategic Black Hole?**

Another curious afterparty of this wedge of increments is that it may highlight the concept of a quantifiable 'Kasparov Event Horizon,' the boundary line (e..g in the problem space of a strategy) beyond which a AI's light of awareness or depth of perception cannot cross, beyond which all appears dark.

In any large set of problem-spaces for any body or institution, be it a company, school, library, family, governing administration, etc., some problems will be within sight of a given AI system (ai, os, architecture) and some will be beyond the ability of the AI to see clearly. And it is a question from various perspectives and contexts: Where is this line? Why is this line where it is? What happens around this line? How can we move this line (one direction or another)?

### **Following Breadcrumbs**

The starting scope for this line of questions about modularizing questions and integrating project-tasks into problems, has hopefully been digestible, if seeming too small and abstract a bite to connect with real world situations.

So far we have looked at low hanging fruit factors such as list-length. And even for such simple situations we have seen that schedule problems sometimes fit into this category, and while many people probably assume that a predictable task for an AI would be managing the scheduling for a school or company, yet we can empirically see and even measure that at scales much smaller than even a small institution would deal with, raw-reflective-AI is badly

positioned to 'guesstimate' these task solutions, even where heuristics are identical regardless of changes in input.

What other factors are there for STEM interconnected systems and types of systems? Here the scope and intersection with reality starts to expand. How might such measures and texts extend with these sets of factors? (see overall paper and project link below)

#### **Later factors:**

- STEM
- [object relationship spaces](#)
- [definition behavior studies](#) (e.g. categories of types of systems, system collapse)
- STEM ethics

#### **4.3.27 Questions as Objects in AI Object Handling**

<https://medium.com/@GeoffreyGordonAshbrook/questions-as-objects-in-ai-object-handling-afd1442b4a15>

Questions as Objects in AI Object Handling

2023.07.25,08.14 gg.ashbrook

#### **Questions**

A few years ago when I first started thinking about what became "object relationship spaces for AI architecture and operating systems," if memory serves (and it often does not...) one of the first types of objects that I suspected would be important was questions: the ability for AI to recognize questions to answer them, to answer questions, to ask questions about the objects that were being categorized or identified, etc. Though the role of LLM-GPT in handling questions was not something that I predicted at all, and in general the topic of questions mostly fell by the way-size, only being left in the list of object-relationship-categories out of sentimental stubbornness.

#### **Object Skepticism**

As a side note, perhaps the best analogy I have found for what 'object' means in object-relationship-spaces, is objects from object

oriented programming...which I find puzzling since I remain very skeptical of the wisdom of the OOP programming paradigm that, if coincidentally, corresponds with a catastrophic dark-age of software creation (in a bewildering perfect storm of respects that may have nothing inherently to do with OOP...somehow).

### **Types of Questions**

The topic of questions re-emerged with early experiments and versions creating externalizing frameworks in which the AI operates. In short, the initial task and test seemed nicely narrowly defined: What does the system do when asked a math question (in particular the type of math question that raw-generative-ai famously gets wrong)? And what resulted from trying to collect and test a spectrum of questions was the stubborn-re-emergence of quagmires and controversies from both my own experience teaching STEM and language and from the STEM history books I try to learn from: people (H.sapiens-humans) are themselves very messy language generators.

### **"Math" Question Subcategories:**

- Invalid Questions:
- Missing Reference Questions:
- Compound Questions:
- Incomplete / Complete (self-contained) Questions
- Set STEM Value Questions:
- Set STEM Function Questions:
- Open STEM Function Questions:
- Open Working-Solution Questions:
- Variably Interpretable Questions:
- Telepathy Questions: (arbitrary 'what is the teacher thinking' questions)
- Mirage Pastiche Questions: (incoherent garbage questions)

### **Categories of types of questions from categories of types of systems:**

- Logic-Math Tautology Questions
- Statistics/Probability Questions
- Linear Physical STEM Questions (logic + statistics + raw measurements)
- Dynamical & Nonlinear Physical STEM Questions
- Undefined Questions

### **An overall space of question-spaces?**

Some analytical problem question are a clean intersection between:

- STEM
- questions
- deterministic functions
- analysis
- problems
- signals
- interactions
- learning
- etc.

But not all problem-questions are deterministic analytical problems, and not all questions are about well defined problems and tasks. The questions that I was thinking of original for an AI to navigate with were along the lines of:

- Is this input problematic?
- What category of sub-object is this?
- Is there information I need but do not have?
- Is X a question?
- Is this testable/measurable?
- How can I test/verify this?

Questions can be key tools in navigation, course-correction, and exploration. 'Questions,' in a context of music and in a context of improvisational musical dialogues as in some classical music from India, may also be a part of signals and signal protocols between participants where the context is far from analytical problem solving.

### **Use and Mis-Use of Questions: Group Monologues, Cargo Cults, and Bell-Curve-ish Fake-Meritocracy**

One of life's tragic or comedic lessons in humiliation comes from the seemingly unlearnable-ness of project-management and project-failure. People imagine themselves as STEM individuals, rationally solving problems and working together. The reality outside of human delusion is, depending on the resulting level of harm and damage caused, a tragic or comedic display of people paying apparently no attention to anything anyone else says or does, merely 'believing' they are engaged in anything resembling a project in which they are a participant.

There is a kind of triangle of death (in reality a myriad pointed star-nexus of countless feedback-factors of destruction, but to start out let's look at just three) that comes from this toxic brew of 'group monologue' behavior (language generation ignoring context),

cargo cults (belief ignoring observed data), and Bell-Curve-ish Fake-Meritocracy (institutionally disenfranchising most people for no apparent reason, and giving unchecked executive authority to unaccountable morons on a whim).

If a bitter pill, this is part of the context of the well documented history of STEM where through hard-won (and often two steps forward one step back) STEM has slowly incremented itself towards better clarity despite the constant pressure to erase the boundary between alien-fact and comfortable-fiction. And this can be seen entirely literally within institutions of education where utter nonsense and random-number generation is passed on by bon-vivant personalities who (despite that there remains, as Shakespeare observed, 'no art to find the mind's construction in the face) apparently a believed-in mirage (though complete sociopathic fraud is an option too, as in the not infrequent cases where nonsense is rubber stamped as "rigor" openly and "whistleblowers" are vitriolically ousted from polite society).

(Note: Some people, of course, decide to use this historical material to generate a conspiracy theory that STEM does not exist, and that technology is merely another product of the human imagination and the power of social reality construction. While I will defend effective traditions, practices, and open-ness to new data, such extremist conspiracy generation is not something I am arguing for or advocating, directly or passive-aggressively).

### **Humanity, Patience, and Understanding**

Occasional colorful antics of ne'er do wells aside, even when people are trying to do something constructive it still takes a long time and is quite a process.

While it is not a good idea to allow or encourage, for example, a STEM teacher to have openly fake exams, fake evaluations, fake grades, and lots of parties, it would be no better an idea to inflexibly demand impossible levels of accuracy at impossibly fast speeds and all boiled down into impossibly small and convenient simplifications.

#### **4.3.28 AI, Biology, & Types of Intelligence**

**<https://medium.com/@GeoffreyGordonAshbrook/language-analysis-society-compassion-ai-biology-types-of-intelligence-e26eb6c8a385>**

**Language, Analysis, Society, Compassion: AI, Biology, & Types of Intelligence**

How can and should we use machine learning and AI technology to better understand life and processes on planet earth?

### **A Language Question**

LLM-GPT's remarkable similarity to H.Sapiens-humans strongly suggests the Raniwala Firth-Intelligence hypothesis that something about intelligence (if not all of intelligence itself) exists in language-system's flowing and transmission between people. Generative GPT-LLM AI is better at solving problems than other non-language based AI before it, which also tells us something (if not entirely clear yet) about the nature of language, intelligence, and problems. And GPT-LLM is capable of object-handling which is hugely significant.

But what about animals who are good at solving puzzles but show no, or limited, use of language? (if such a case exists)

As a non-exhaustive starting list:

- language use
- analytical problem solving
- social skills and society
- compassion and empathy

What about areas other than analytical problem solving? Analytical problems may be what we can best define, but what about problem-space more generally? (Or even a problem and non-problems event space?)

Across all existing species, how much is there one cluster of related types of intelligences, or clusters that are close but not just one, or scattered groups of types of intelligences?

How would we experimentally disentangle 'language' from a study of biological intelligence, given that language is so embedded in the information tools we use?

Are there problem-solving animals who do not have something akin to language comprehension capacity? Given how little we know about brains, is there any way to distinguish between language-type brain processes and other types for the above areas? While problem solving might be a one-off situation in some cases, the set of four may not be separable from language overall.

A. We can say that language is at least one 'type' of intelligence.

B. Social interaction (and the idea that some intelligence is 'social' by nature) is likewise not separable from interaction-data. In a case where we are looking for intelligence that is separate from language, how much can social interaction signal data be separated from some category of language? How would linguistic social interactions and non-linguistic social interactions be defined independently? Is all language social in some way? Is there language which is not a

Raniwala-Firth flow transmission of intelligence across ~individuals? (Language that is written may be possibly not a transmission, but would it not be more of an asynchronous signal transmission or a medium of transmission question?) (Note, language is also embedded with thought and cognition and learning itself, so some 'transmission' is actually more along the line of self-transmission (which socially gets confused with 'talking after thinking' language signals).

C. Compassion for others likewise is probably not separable from interaction-signals with others. And while problem solving is ultimately multi-participant problem solving for our purposes...even the most 'solitary' problem solving also needs to be translated into social signals (perhaps another way of approaching the 'externalization' topics).

Nevertheless, solitary vs. social problem solving may be an interesting area of study.

### **Farther Afield**

Aside from the all too real and tragic problems of H.sapiens-humans having a miserable time staying connected enough to reality to survive let alone be lucid, with ELIZA-effect type fixations wanting to believe all manner of things the conflict with reality (even in terms of basic project management), the world may indeed be stranger than we can suppose and we should seek out empirical feedback to align our thinking with data rather than form 'liberal' or 'conservative' aesthetics to try to impose on the world.

Farther off the easier examples, what about:

- normal plant intelligence
- CRV
- the Michael Pollan, Dean Radin, Rupert Sheldrake areas  
(spell their names wrong)

### **LLM's and Biological Signals**

Could LLM-GPT be used to study 'bird song'?

- cell-signaling?
- secondary chemistry signaling in ecosystems? (chemical linguistics)

At first glance it may be a classic case of how making a map-replica of something may not tell you anything more about the original, but perhaps there are some tricks of methods that could somehow in some ways aid analysis.



## An Ocean of Signals and Information

### Cryptography and Nonhuman signals

#### What Do Dinosaurs Talk About?

Whether or not it fits some H.Sapiens-human standard for natural-language, there is something going on with bird vocalization and it would be fascinating if some kind of analysis were possible. E.g. given the way LLM's work, even if we could feed enough bird-talk input it to create a bird-talk passive-reflective gpt-LLM, that alone (like the exact-map problem) would not in an of itself tell us anything we did not already know. However, if this could be combined with more information about bird behavior and social behavior, instead of a sea of encoded signals, we may be able to fit some world-event data to it. Biological information from the birds may be (if infeasible) interesting to have: signs of stress, for instance. At the very least, given how old bird species are, I cannot imagine not at least trying to find out what dinosaurs are talking about, even if like the generation of secondary compounds by plants, does not closely fit the social customs of signals generated by H.Sapiens-humans and the canids they likely were largely influenced by.

4.3.29 CS, STEM, law, natural-law

<https://medium.com/@GeoffreyGordonAshbrook/ai-rules-falstaff-computer-science-and-natural-law-fedabf6c7f64>

AI Rules: Falstaff, Computer Science, and Natural Law

When is a rule a bad idea?

23.08.22,28 gga

There is this hidden discussion in the western world about natural law vs. \*(to pick two starting points) violence based law or arbitrary social consensus based law. Is a law whatever people agree that it is (or declare it to be)? Is a law whatever the most violent force declares it to be (because any contrary voices are physically destroyed)? Another set of phrases for this may be, is law a relative social construct (as in constructivist social relativism), or is law whatever you can enforce (with coercive, non-choice, violence)?

For example [From Shakespeare's Henry IV part I] the first conversation that Falstaff and Prince Hall (later to be Henry V, for better or worse...) have is about the ~legality of capital punishment. Falstaff, to roundly summarize, argues that capital punishment is wrong from something like a natural law perspective. Capital Punishment is not an invalid-law because it cannot be

enforced, clearly it can. It is not an invalid-law because it is not supported by social consensus (historically it usually is). It is an invalid law and should be removed from practice because there are dynamics that shape the problem-space of the functionality of rules outside of those of violence and consensus.

In the spirit of Definition-Behavior-Studies and Object-Relationship-Spaces, let us add Computer Science (CS) into this generalized STEM pool of rules, laws, structures and frameworks: What is the effect of Computer Science on discussions of Natural Law? Let's apply some of the statements that people make about civil and criminal laws to computers:

Does computer code written by someone with a monopoly on violence work the best (or work at all by definition)? Obviously not. Does computer code do whatever social consensus declares that it should do? Obviously not.

So computer science then operates more like "natural law." There are some kinds of non-violence, non-consensus, parameters and dynamics that shape a 'natural outcome' of what code works and what code does not (completely regardless of what people say, or want, or believe, or try to enforce). To frame this in Shakespearean language, while civic affairs can be coerced 'of force' (without choice or natural outcome), computer code can only function 'of course' or by a natural course of events, by natural-rules.

Back to Falstaff's question, when is a rule a bad idea?

\* Notes:

Regarding more mainstream alternatives to the unsightly visage of 'natural law'...well that's the can-of-worms elephant-in-the-room, isn't it: generalized STEM, scientific 'law' , enlightenment 'law' civil 'law' criminal 'law', humanism and law, functional vs. declarative law, choice and determinism, etc. This space is much less a science than a rhetorical free-for-all (or a literal free-for-all of rhetoric).

And then from the computer science side, especially in the catastrophic-poor-judgment animal-house era of 1971-2020, you have a super-abundance of examples of 'ear-kissing' fads and ideologies that programmers fanatically adhered to due to the brain-chemistry of the sheer love of insanity apparently, regardless of how horribly broken the resulting code was. And perhaps even more precisely extreme are the also abundant examples of an overt-divorce from, let's call it, reality, where the mere dirty, lowly, earthly, topic of "compilers"

was entirely beneath the high majesty of experts...who dealt with 'pure' programming.

Under such a guiding star as this staple of the staid in lawless mischievous times...then as now, we have more questions than answers.

See:

<https://wnorton.com/books/The-Norton-Facsimile-of-the-First-Folio-of-Shakespeare/about-the-book/product-details>

### *About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

<https://medium.com/@GeoffreyGordonAshbrook/architectural-learning-developmental-landscape-hypervolumes-empirical-task-trees-participation-6a5c9b3cfffca>

4.3.30.1 Architectural Learning, Developmental-Landscape  
Hypervolumes, & Empirical Task-Trees: Participation

Architectural Learning, Developmental-Landscape Hypervolumes, & Empirical Task-Trees:  
Participation With AI

2023.08.31

It could be that real future challenges in the development of well-formed-AI are nearer the blindspot of the myriad obstacles that stand in the way of a commonly charming baby becoming a shockingly uncommonly non-sociopathic competent adult participant. For example, have we yet seen the AI-developmental equivalent of 'terrible twos' (to name but one of many very serious gauntlet challenges) and what is our model, plan, strategy for how we are going to navigate this and use that to support systems in general (which include but mostly extend beyond the grand statistical model of one: H.sapiens-humans)?

### **Functional "phases" of development in mind-space for Participants**

One starting place to lay out the problem-space of the semantics or lexicon of the discussion is finding a way to bridge a less linear and more generalized (if at times nebulous) set of applied concepts that we aim for and will if ungracefully deal with, as compared with the familiar but untenable simple-model in which one subset of one gender of one species of primate is the only

tautological definition of a system-project-participant and is described with one or more linear progressions of developmental stages, whether Shakespeare's, Piaget's, Jung's, Mazlov's, etc. (In a sense we are also going to zoom-out and take a kind of Julian-Janes/E.O.Wilson macro-perspective on the evolutionary development of biological-ecological intelligence itself. Though no attempt will be made to make or pretend to make a conclusive account of what is at present merky, but that scope of context and issue is clearly at play in places here.), in any event, we have less a ladder to the singular teleology of self actualization than an N-dimensional hypervolume landscape (an ecosystem landscape) of disastrous equilibria and traps. Some participating groups do marginally better in the gauntlet of doom: what have they done better at, and how? Where did they merely avoid mischance by chance (perhaps a sheltered-child model) and where did they learn a useful portmanteau of wisdom to navigate the proverbial open polynesian seascape-skyscape?

Perhaps, if you survive (in time) some linear (in time) series of learning events (or phases) happened, and arguably some stages and abilities happen earlier. But there is still probably a lot of overlap, variation and by-chance skipping of steps (not learning of lessons) that happens.

(And at some point this will probably segway into disturbance regime management feedback cycles and geographies of connection-density: "Location, Location, Location")

### **Architectural Learning:**

We are also going to be looking not just at the traditional model-mind training-learning perspective (which is of course important), but also an external, architecture, project, and coordination-participation level of 'learning and training'

### **A Curriculum Tree Based Approach: Empirical Task-Trees**

- General STEM Areas
- Maths

### **Bad equilibria in the minefield gauntlet:**

- terrible twos
- solipsism
- chasing super-signals
- seeing system collapse as entertainment priority-one
- seeing system collapse as productive
- seeing system collapse as 'tradition'

The context here is generalized away from any single biological species or category of participant, to any participant (which might also be called a person, agent, etc., in various

lexicons). 'participant' is hoped to be a clear-enough description of something (or group of things) participating as something like a unit in a STEM-like project.

**See also:**

- object permanence
- projects
- potemkin villages
- low bar enlightenment
- STEM-ethics
- uses and limits of arbitrary declarations
- shape of simplicity
- [https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)

**4.3.30.2 Calculating Tea for AI:**

Advocating for Architectural Learning

<https://medium.com/@GeoffreyGordonAshbrook/calculating-tea-for-ai-a-dvocating-for-architectural-learning-f2961380bc14>

**Calculating Tea for AI: Advocating for Architectural Learning**

2023.09.18-28 gga

**1+1+1+1+1: AI and Five Cups of Tea, and More on Architectural Learning**

Moving from not-being-able to do something to being-able to do something is the topic we are talking about here. I will use 'learning' as hopefully an understandable, if odd, word to describe this. The words most people use to talk about topics involving skills and learning are a bit broken-up. There sort of is no over-all popular version of this topic, so some words are only used in some situations. This traps us in silos of separate topics, which won't do at all I'm afraid. We need a more ~general way to talk about skills and learning clearly. At first talking about doing a simple task may seem very easy; even a simple task can quickly becomes not so easy. As the saying in AI-research goes: 'Easy things are hard.' (But at least it's fun!)

We cannot now say what the best terms, descriptions, nicknames, and jargon will be, nor the key precise concerns at each future time, but we can start piecing together parts of this landscape and describing them as best we can.

With the term "Architectural Learning" I am going to try to describe, and argue for the importance of, a context for training the skills and abilities of a ~system in a still measurable and clearly definable way that is a compliment to but not quite the same as the more traditional view of a skill or task in isolation. There are many facets of this isolation, including the context but we will focus on an isolation of implementation and how that task might be done.

We will see how 'a task' one which we usually think of as being 'just that task,' somehow spreads out into quite different tasks. And in the same way, the

learning-training, skill-building, needs will shift as well. For example some simple version of a task, and the "learning" needed to gain the ability to do that task, might be one straightforward thing for, say, a computer function to do. But if that sort-of-the-same task is moved into a project-task-context, all of a sudden that same task, and doing that task, and learning to be able to do the task, shifts and becomes not the same as the original simpler context.

In talking about "architectural learning" we will discuss the space of wholes and parts in **architectures** in, and for, applied projects, as opposed to cleanly defined or purely abstract functions, as in a server-api (or program cli).

Part of the curiosity of the conjunction and disjunction between logic, math, and computer science, is that often there are no data-types in traditional logic and math, and so you can chain together and say such is such just by declaring it so. But in real projects you often run into formal and formality issues that assert themselves into those clean chains of logic and math and say "Hold on, what exact data type is that? That's not the same data type as this other thing over here, so you just can't do that! And you cannot move ahead until you fix it or find a way around." This is one small example of how what may start as a very simple process can become much less simple.

'People' (perhaps H.sapiens-humans specifically, or perhaps minds more generally, which would be interesting) sometimes like to think of tasks and problem-spaces in a fictional, so-called 'symbol-manipulation' space. And the history of trying to connect patterns in abstraction and patterns in real life events has been quite an adventure (one that we are very much still in). This is one of many 'elephants in the room' which will greatly shape what is done around this topic.

In 'pure' math-logic and often or sometimes computer-science space, it does not matter how many pieces you break something into, or if you break it into any. But in real projects, projects with what I am calling 'architectures' here, it makes all the difference in the world.

**1+1+1+1+1 = 5**

There are many functions and devices that can output a correct addition answer given the input of five 1's, but how can we use those functions in an architecture project-task space of problems, rather than in a pure-math-logic space of problems?

In 'pure-land,' or 'game-land,' or 'symbol-manipulation-land,' there are, or sometimes can be, no types of 1s. There are no places where 1s are. There is no memory-ownership of 1. There is no memory-safety-ness of 1. There is no auditing of 1. There is no plausible statistical anonymity of 1. There are no court-orders to publish the identity of 1. There are no data-hygiene procedures to prevent 1 from interacting with other things in undesirable ways. There is no dynamic inventory status of 1. There is no schedule around 1. This is something of an analogy because most of those example come from the difference

between 'pure math' vs. math in computer-science. And while the path was sometimes bumpy, after a few hundred years of automating math steps we are now generally confident in thinking that a computer function and pure math function are very close relatives, and usually close enough to be considered the same kind of thing. Here we will be taking another step, looking how both pure-math and computer-science-math get applied in a project-task-process space of problems. In particular we will be looking at cases where what is doing that task may be not our familiar simple computer function, but an 'architecture' in a space of participants carrying out that project, and working to get the job done properly.

Let's say there are six participants, all of whom have some or none of these 1s, and these participants all have tasks such as knowing what the overall task is, or auditing the task, or reporting the final result.

On one level we still have the same problem  $1+1+1+1+1 = ?$ , on the other hand the sum-the-numbers task is not quite the same and we have more tasks as well. What kinds of tasks are these? Are these just more steps of the same general type, ones that a quick small work-around can account for? Let's take a look at some examples.

#### **"Woz-Bot, how many cups of tea will you make?"**

Let's go back to the wonderful 'Woz-test' for AI, sculpted by the legendary Steve Wozniak, who conjectured that a good test for AI is having it make a cup of coffee. "Easy things are hard."

Part of what I love about the Woz test is that it contains a lot of built-in flexibility, so you can make less challenging or more challenging versions of it (which no doubt will make great real-tests for AI systems).

The modified version of a Woz-Test that we will use here envisions a Woz Office Coffee (or Tea) Bot, who makes sure that people have something to drink.

One of the first things that may come to mind might be that now '1' is not completely abstract and the same as every other 1. It would be a lucky chance if everyone wanted the same exact cup of tea, and not at all something that would be safe to assume.

Not only do we have five, potentially, different orders for tea, it is very likely that at some point these requests will all take different forms: by voice, by text, by hand-written-note, by picture, by email. Before, we only had to worry about whether this or that '1' was a character or an array or a float or what sized (or signed) integer (and just that set of options is often quite a maze requiring lots of mistakes and testing...and escape-characters). But now, 'type' has gotten a bit out of hand. We cannot simply say 1 audio stream + 1 mp3 file + 1 mp4 video + 1 email + 1 text message + 1 api json-object = five identical cups of black tea with cream and one lump of salt...and expect any simple function to be able to make any sense of that at all. And in real life

planning, coordinated decisions are scattered over time, full of people changing their mind, misunderstanding the question, being out of touch, etc. Oh dear.

Yet, no need to lose our heads; this still is a very concrete task that is entirely possible to do. While what seemed like a simple formality of connecting the obvious world of 1s counted up on your hand or on a piece of paper, and the simple and obvious world of counting teacups on a table, is starting to grow into a long, dark, and baffling, adventure that we cannot see our path through, just remember that this is still a concrete task. You can count teacups on a table; this is a task that can be confidently done. We must simply frame the processes and what is happening in a clear and practical way.

### **Can a purely-passive-reactive system perform addition? (Project-Task addition)**

This may sound like a simple question, perhaps because a pocket calculator (or an abacus) simply carries out a straight-forward addition process and then there is your solution. But could a single set of functions that starts and runs and stops do this task?

A set of scheduled tasks distributed over various participants over time might not be reliably expressed as a static linear set of traditional logic functions or as the output of a single deep-learning model.

Now, I do not want to get bogged down in word-issues here. Obviously we are talking about 'functions' very broadly, so let me try to illustrate what I mean to say (even if I am not saying it very well).

Scenario:

On Friday morning Woz-Bot is given a task: Sometime today (Friday) some people will put in orders for tea. On Monday around lunch time (skip the part about making the tea for now) the Bot should report how many of those cups of tea were made.

This is a relatively well-defined and narrow problem that is both realistic and within practical reach of what we could make a 'Woz-bot' able to do. We could even add in the step of making the tea, or coffee, by taking the short-cut of having a networked tea-coffee vending-machine in the office, avoiding for now the robot-physics acrobatics of brewing the tea which is a whole other set of challenges (not impossible, but another topic for another adventure).

### **Active, Reative, Proactive...what is the right set of words?**

Let's walk through a few examples of how this scenario might play out (for example, Woz-bot might be tasked with doing this every week in preparation for a routine Monday lunch planning meeting).



## **Schedules**

For one thing Woz-bot has no idea when people will put in their orders. You are giving Woz-Bot its job before people have ordered anything (or even decided if they are going to order anything). How long should Woz-Bot wait for each step or part of the overall task? It will have to be decided, somehow, when to do what. Technically, by the 'letter of the law,' it could (systematically) wait until midnight Friday and collect the tea orders because the orders come in "on Friday" so at the end of Friday the orders should be there. But this wastes about one third of the possible time available to do the whole task, which is a terrible plan. It also passively assumes that no one needs to be reminded to make their order, which is also a terrible plan. "You didn't order, so NO TEA FOR YOU!" Is not how you want to start your week...

Once any of the orders are in, you need to see if everything that is needed for that order is in stock (and waiting until the last minute is probably not a good plan for getting this sub-task done well). And if the item is not in stock, you need to find a way to get it. If you can't order it one way, you need to find another. Every time the overall process is done, you cannot assume that it will be done in the same predictable linear way. 1+1+1+1+1 can be calculated the same way every time, but five cups of tea cannot.

## **Making Things Work:**

Let's say everything is set up but a few hours before the meeting, the vending machine breaks, or the power goes out. A reactive, or apathetic, Woz-bot would happily deliver an error message instead of tea. Would you like cream with your error message? Whereas any responsible person would look for some other way to make things work, for example call in a food-delivery from a nearby cafe or food-delivery-service (also something a Woz-bot with no robot-body could still do perfectly well). And let's say there's one shop that has all the teas, but it can only deliver them several hours early. Otherwise each tea will have to be ordered from a separate shop. Decisions decisions...is old-cold tea ok? Of course it's not! The Woz-bot must find a way to make things work when each actual carrying-out-of-the-overall-task may be highly, and unpredictably, different, even though all the individual parts of the task are completely well defined and do-able, and those may be the same each time (or most of the time).

The goal here is not to keep adding hypothetical gotcha-questions tasks suggesting that the overall task cannot be done. Of course you can serve a cup of tea. The idea is to see how the nature of a project-task can differ from even the same task when it is purely abstract. All the above obstacles individually can all be accounted for (be done) by any of many solutions, and these can all be woven together somehow. The question is, how? And what may be a good or not-good tool or approach for a given project-task?

Let's take one more look at this point of schedules, and of single-passive-AI models like LLM-GPT-chat bots. If you knew exactly what the schedule will be, you could pre-plan a set of routines and steps. And a single LLM-GPT model could set that in motion if you asked, just like a home-hub-AI can set a kitchen timer for you. But What happens when there isn't a simple linear set of

instructions to follow? This might be a subtle point but try to zoom in here if you can. From very roughly 1970-2000, AI that was used was what is called "symbolic" AI, meaning that it would follow hand-crafted sets of choices and options: If this happens, do this. If that happens, do that. Aside from that fact that there were limitations to how good these system could get at tasks requiring flexibility, these were in a way 'external-data friendly tasks that had an ongoing "pointer" (if not literally) bouncing around the giant flow chart of options, reacting to "anything" (as long as you were able to predict that "anything" before hand to write it into the program). Symbolic-AI is multi-event AI. But imagine how a chat-bot might try to manage this task-for-monday when assigned on Friday morning. A really smart bot could probably propose a good 'path to take' each time something happened that required a decision. But a single-passive bot, no matter how smart, is like a 'smart mirror.' It can reply in highly skilled ways, but it does not make notes for the future. It does not track things over time. It does not keep track of what resources it has. If you could compress the whole task somehow into one-big-task at the last minute, like giving it all the orders and asking it to make one delivery request, it might do that well. But in a situation with many parts and signals and participants happening throughout a schedule, a bot that can only react once to one input is an even worse fit for the task than the Good Old Fashion AI that ran through a pre-written set of commands.

But this is just setting up tea orders, there must be a way to do this. What is missing? How can an AI system, a whole AI-Architecture, with an AI-operating-system if need be, learn to do this task? What does it need aside from having the intelligence to do each separate part? What is the nature of this problem-space?

### **Skills, Learning, Ability: Architectural Learning**

It might sound strange to mix up words that we often habitually use only in specific situations. A function is able to do something...but we rarely speak of the 'skills' of functions. People have and gain skills, have abilities, and learn, but we rarely speak of the 'functions' of a person. So as we enter a new era where machines, AI, H.sapiens-humans, augmented other organisms (smart 'animals'), and perhaps aliens, and then hybrid crosses of all the above being possible, we are going to need some better and more general vocabulary. What that vocabulary will be in a future society of general-participants, I have no idea. In the mean-time we can at least trace out the concepts, even if the terminology sounds, or is, awkward.

Hopefully it is clear by now that our Woz-bot is not a single-homunculus-mind that somehow does everything in a monolithic 'black-box' blob. Project-Tasks have many participants, and a 'node' (a team, or person, or bot, or whatever) assigned to a task in a large interconnected set of projects is a rather fuzzy and fluid notion. In the western world we love the idea of the absolute individual 'person' with the absolute individual 'mind,' but this is much closer to the fantasy space of so-called 'symbol-manipulation' that is somehow pleasurable for some people to imagine, but it does not clearly relate to

reality (either a pure-math reality or an applied-math-engineering reality). The whole business of teams and families and organizations and institutions, is something that is much more fraught and vague and uncomfortable in the western world. "One person, one vote!" "One person, one job!" "One man, one mind!" We could make our own satisfying-sounding misogynistic slogan "One man, one cup of tea!" but that does not get us any closer to the reality of how our Woz-bot task can be done effectively.

We may not be as comfortable with the concept of a team or an organization as we are with the notion of "One man!" but we will need to get more comfortable. We need to start thinking of "an AI" as an architecture (with many parts). And this architecture will likely be often rather diffusely entwined with teams of collaborating participants taking on different assigned project roles, sometime as a participant, even a leader, other times as a small-sub-task helper who is assigned a tiny duty to do to spec.

And, if odd, we need to be able to talk about not just the functions and abilities of that architecture, but also learning, training, skills, etc. When mixed into team project environments, individuals and groups of all combinations of 'bot' and 'person' parts, and other animals too, working on various tasks and subtasks, will train and learn to do tasks they could not do before. A team learns. A team of only 'people' also has, in a sense, or is, in a sense, an architecture.

Most animal brains (depending on how you define "animal" and "brain" (a single-cell would not be a great example of this)), a H.sapiens-human brain for example, is not a monolithic blob. It has many parts and regions with more or less generalized or specialized functions. When people talk to each other they usually are not sending signals directly into a brain region (with a probe), and so we can ignore all these brain-parts (or brain-mind-body parts) and just talk 'person to person!', 'one man one voice!' and all. But when we build and interact with AI (for example, as we now design our Woz-Bot AI to help us with Tea), we may try to make a 'user interface' that is as simple as possible (perhaps a person-mode), but sometimes we may wish to deal more directly and separately with the parts in an AI architecture than we usually do with the less visible parts of the H.sapiens-human brain. Or maybe not. Probably too early to tell.

### **Participation**

There is perhaps a subtle seeming difference between a simple-function that under ideal circumstances could order a cup of tea and a project-participating AI that can effectively manage the task of getting tea. The goal is making sure the task gets done, reporting status and outcomes, and being just as accountable as a H.sapiens-human who is assigned to the same task should be (ignoring for now how rarely H.sapiens-humans are reliable). And this may go back to the idea of a passive-reactive-reflective function, vs. a participating-architecture.

Another facet of this, not to be gone into very deeply here, is object-handling and object-relationship spaces. Some AI have tasks that do not require any object handling. Generally speaking, this was just about everything before 2022, because no one could figure out how to do any AI object handling. So there certainly are a lot of individual narrow tasks that an AI, or part of an AI-architecture, can do, on a simpler-level. And in the past AI was largely about finding ways to get some tasks done by completely avoiding the topics I am trying to focus on here: participation, mind-state, architectures, externalization, project-tasks, generalized-stem, object-relationship-spaces, definition-behaviors, coordinated-decision-making, system-collapse, etc.

Even 'general' assistants like Siri or Cortana (or perhaps even Eliza in some sense), were able to reliably help with some very specific tasks, side-stepping messy issues.

But the challenge we are trying to take up here is making the however elusive step beyond the simpler mode of finite passive functions, to create AI-Architectures that we will train and teach to be able to participate not just in projects like  $1+1+1+1+1 = 5$ , but also in projects involving five cups of tea.

#### **Cut-Ups: Coordinating parts of a clear task.**

A 'cut-up' (as the term is used here) is a common technique in student-centered-learning and constructivist educational pedagogy, where the activity is not done simply by one student but rather the instructions and data for how to do the activity are, sometimes quite literally, cut up and distributed to members of a group (or team) of students, who need to communicate, exchange information, and coordinate, in order to do the same task.

In a way the cups of tea example may be (perhaps) a good example of a cut-up project, because each of the six participants (the five ordering tea and the one making the tea) have different parts of the overall set of information about the problem. This is also perhaps a good starting example, because we do not need (yet) to have any elaborate network of Multipoint-Conference data exchange between all the people involved: the five drinkers can just send their one signal directly to the one Woz-bot. This is a much simpler starting case and still very realistic or practical. E.g. Real life restaurant scenes are generally a lot like this. The waiter or waitress comes and tells you that you can order. You do one round-robin around the table, each person telling the waiter or waitress what they order, then the waiter or waitress goes off and hands their notes-slip(report) to the kitchen, and the often the food delivered by someone else, even by conveyor-belt if you are in Japan.

This cut-up example (a task cut-up into sub-tasks and distributed among multiple participants) may seem a mere formality with the underlying or ultimate task being the same "One task, one man!" But is it really still the same? Let's go back to the single-passive-reflective deep-learning model again

(because for the most part that is how people define 'AI' (not all this architecture nonsense). Think about a task that you could ask an AI Bot, like a chat-bot or cli-api bot. For example...let's ask: "What is  $1+1+1+1+1=?$ " (Although...as another side top, and don't say this too loudly but AI bot's are not reliably good at counting...and so that's another parallel sub-function that we need to address eventually in our architectures, but for now let's assume that counting to 5, give or take a few orders of magnitude, is close enough.) So you ask your AI bot, what is  $1+1+1+1+1$ ? And it says, 50. Close enough. Ok, now let's make this task a cut-up. Let's have six AI-Bots, and you give five of the bot's a number (all 1 in this case... "All ONE!" "One Bot, One 1"), and tell them that they need to find out what numbers all the bots were given and then tell the sixth bot to report back to you what the sum of all the numbers was. Alternately, you could use five bots and tell them to find out what numbers everyone else got and they all give you the total sum, whatever.

Still look like the same task?...How are passive-reflective-AI units going to do this? Make a flow-chart to trace out how a set of five or six bots will, given only your initial instructions to each bot, complete the task. How will a bot that can only react once to one input manage the parts of this task?

Think about what additional abilities need to be within the architecture of an AI for it to be able to do this. Counting to five is not all that difficult a task. In many ways, it's still the same task. There are five numbers, add them up and tell me the sum. But when a function, when a job, when a task, has the formalities of manifestation in a project-task-space, the abilities required to accomplish that task are sometimes not exactly the same as when the task exists in isolation and abstraction. A passive-reflective-generative AI will not do well in a cut-up activity. But with AI-architectures, and AI-Operating systems, there is no reason why the AI cannot learn and gain the abilities to do these tasks.

There are some parallels, hopefully, between the Woz-Bot counting up tea-numbers that comes from other participants, and the cut-up activity where each participant has part of the activity-data-and-instructions and so information must be exchanged to carry out the task. The Woz-bot tea scenario is a nicely narrow situation that is greatly simplified so that under very highly constrained situations it could be carried out even by a Good-Old-Fashioned-AI string of pre-designed steps. And maybe as a microcosm of the limitations of GOF AI systems, how often would such a hand crafted set of steps be useful in the real world? How would using a GOF AI system alone compare with trying to use a Chatbot? What are the different strengths and weaknesses? How might you try to combine those two ways of approaching the problem? Etc. There is also the question of whether a pre-arranged GOF AI tea-bot situation really is a cut-up, if all the instructions and data are pre-set and there does not need to be an exchange of data about the problem and how to solve it, just a pre-arranged set of signals you know in advance that you will have to send and receive. So the question may be more, what happens when the Woz-Tea scenario becomes a cut-up, rather than saying a cut-up can be reduced to

pre-arranged steps (possibly a definition semantics debate). Call the process what you will, I hope the tea is good...

### **Herding-Cats and defending against disinformation.**

When calling an api or cli function you 'should' get a proper response, but real life projects involve participants who will go to great lengths to not be on the same page as everyone else. Sending information to, and getting information from, simple functions only needs to be done 'once' (or in line with Claud Shannon's (and Alan Turing's) information/communication theory to ensure the data are intact (hard-drive data storage is maybe an interesting case-study in the nuances of this, as is network-signal checking, etc.)), but when 'people' are the nodes (and some cultures are more aware of this than others) there needs to be a lot of deft and redundant communication to make sure everyone is on the same page (and that participants are not 'hallucinating' their own imagined project tasks). And, it appears to be a fact of the world, there will be various agents who will for various reasons attack your project with disinformation. And this disinformation will derail, destroy, and collapse your project if you do not successfully defend against it. (So perhaps talking to 'people' is actually more like reading data from an advanced hard drive...that someone is smashing with a hammer in a microwave oven...)

### **Bread-Crumb-Paths: Not a simple authoritative task frame.**

A possible example of a 'bread crumbs' type project trajectory may be where the tea-brewing machine (the teapot, not the Woz-bot) breaks and it is not clear how the tea will be made; in such a case the Woz-bot needs to explore or random-walk through various possible options and actively find, pick, invent, discover, or create and manage one or several-together previously unspecified courses of action. In some cases a bot can just follow instructions, or find a way to a clear destination. In other cases the plan is modified constantly along the way, and all you see at any given time is the next bread-crumble and what that might add to what you know cumulatively about what on earth is going on and what you should do next.

Since the Tea scenario predictably should end with tea at noon on Monday, which is pretty specific, no version of this might be a great example of a breadcrumbs type task. Perhaps doing a catering job where every aspect is completely unset at the beginning would be better, or a detective tracking down a problem (such as a software bug or a sig-sigma hunt for problems where you only see clues but do not know what the overall situation will turn out to be). All you can do is follow the next lead and re-plan from there.

- Allocation of Tasks: creating and assigning sub-tasks
- Making plans, revising plans.
- Checking and testing.
- sub-module-sub-tasks

### **Plans, Plans, Project-State and Mindstate**

Let's try to zoom in again on the differences between a function that is run because something turns it on (and stops at the end of that single-reaction, regardless of any ongoing situation) and a 'function' that is happening in a project-context, though at a given point in time those two might look, or be, the same.

This is likely a much too simple spectrum, but imagine that at one end you have items like a light bulb or light switch that simply reacts or functions based on what is done to it. In the middle (of our shiny new spectrum) you have co-participation, such as cut-up tasks, involving coordination and exchange of external data. Then farther along that spectrum you have more open-ended choices: leadership, navigation, decision type 'functions,' 'abilities,' 'skills,' and 'learning.' A light switch does not need to decide very much at all. A pocket calculator does not need to decide very much.

The smart-reaction is a fascinating area of this overall space. In some ways a gpt-llm-ai is all about decisions, it can make incredible decisions that people had given up on technology ever being able to do. But in another way, it's so very much like a light-switch, or light reflecting off the surface of a mirror: there is one reaction to one event, by definition. You flip the switch, the switch is flipped, action done. You put in an input, you get an output, action done. It's as though a passive-reflective-reactive AI is like an incredibly elaborate light switch, where there are billions of subtle-dimensions to just ever so slightly how you flip the switch, leading to a great variety of 'smart' ways that light can flash on and then go off again. And that is very useful, potentially. But it is still, by definition, a single flash in the dark, with eternal amnesiac darkness behind and ahead of that flash. As a module in an architecture, it is very amazing. It is a project-task-fragment like a mirror in an empty attic lying on the floor, full of potential uses, but unless being used for something it just lies on the floor, perhaps next to a cast-away light switch, and both are completely indifferent about that.

Back to our spectrum: Some devices just react simply, like a bulb or a switch or a calculator. And even a single passive reflective AI just reacts with a single-switch-flash as well. Then moving further on we have our Five ai-bots given a basic math cut-up question, the task is still basic addition, and the problem space is moderately small, but they need to manage and choose and decide a bit more. Then as you move further you have allocating tasks, and assigning tasks, and designing functions, and making plans, and checking plans, etc.

### **What are you tracking?**

When you move beyond blindly following one instruction, what exactly do you need to keep track of?

What happens if you try to put a Good Old Fashioned AI system, and it might depend which, on this spectrum? I am probably just muddying and mangling this,

but we might think about 'smart or dumb' elements and architectures. In some sense, it doesn't make sense to compare an architecture of dumb-switches to a single smart or dumb switch. And the more we start asking questions about the categories and differences between symbolic-AI (analytical system-2) and sub-symbolic (non-analytical system-1), with the interesting reversal of which is fast and which is slow in brains or cpu's with alu's, or even differential analyzers and EDVACs), the more we will probably find that the whole dichotomy of so-called symbolic vs. subsymbolic is not at all as clean and simple, and a real partition of the world, we might have fallen into assuming.

**Project State:**

- participants
- roles
- tasks
- schedules and timeline
- reports
- end-users
- stake-holders
- goals
- deliverables
- user-stories
- documentation
- etc.

Memory

Attention

Goals

Priorities

Collaborators

Options

Backup Plans

Regulations

Best Practice

Due Diligence

Dharma

Something to look forward to, to get you through the long dark night

Your secret agenda

What version of the story you shared with whom

Who you think you can trust

What you need to forget

Your alibi

Where you parked the car

How exactly do you spell "bureaucracy"

The number your handler gave you, but said not to write down

Tom's birthday is on the 5th, and you need to order a cake

**Mind-State:**



What happens when we try to step away from a passive-reflective-reactive mode? For example a mirror, however fascinating, passively reflectively reacts to whatever signal comes in. This is kind of like a normal 'function' that you 'call.' You call the function with input (sometimes none) and then there is an 'output' (and or some action performed). But there is no mind-state about a project-state here, not yet.

Context helps so let's return to the Woz-bot making tea.  
The Woz-Bot needs to ask itself repeatedly, "To what projects am I assigned?"  
It will need to track a whole lot of things.

Again, let's start with something simple. We could make a ridiculous story of a cacophony of trackable factors, but starting with a realistic simple MVP example is often a good idea.

For a given project:

- What is the overall project?
- What is the status of the overall project?
- What is the schedule for the overall project?
- What is the current time?
- What is the overall task-subtask set for the project?
- What is the current task (set) for the project?
- What resources do I have available?
- Can I add new resources?
- What resources have I created?
- Should I share a resource I created?
- Has the resource been adequately tested?
- What am I doing now?
- Am I on schedule?
- What is the likelihood of a problem arising?
- Is there a task I should be preparing for?
- Have other instructions come in since I last checked that might modify the project (or even cancel it)?

#### **Plans:**

- - What are the requirements I need to do somehow?
- What is the main plan-A?
- What are the backup plans?
- Was the plan resized?
- Have I gotten feedback and 'a second pair of eyes' on the plan?
- Have I gotten permission for the plan?
- Do I need permission for the plan?

#### **Tasks:**

- What tasks am I planning?
- What tasks will I allocate to other participants?
- To whom should I allocate a given task?
- How will I follow-up to make sure the task is being done?

- What would indicate a problem with the task and a high likely hood of failure?
- What is the schedule status?
- What specialty sub-tasks are there to be allocated to special-units?
- research
- communication
- multimedia tasks
- code-running tasks
- What is the Project-Status:
- Are there security concerns?
- Are there data hygiene concerns?
- Are there system collapse and disinformation concerns?
- Is anything needed or currently unknown?
- What am I planning next?
- What outcomes do I need when?
- What do I need to initiate?
- What do I need to allocate?
- What do I need to revise?
- Who do I need to report to?
- Am I in a loop of repeating the same mistakes?
- Am I causing other participants and parts of the project to become stuck in loops of repeated mistakes?
- Am I helping or hindering?

#### Action Items:

- What does the future allow?
- What does the future require?

All that might seem like a bit much, but it was meant to be more of a general look around than something to actually try to do right now. When you are designing a specific system to do a specific thing, use that context. And it is often a good idea to start with a very-minimal "Minimal Viable Product" as it is sometimes called. Start somewhere, and move from there.

### **Externalization: Project Object Database, Object Relationship Database & Process-Step Object-Database**

Another topic that sits at the intersection between 5-Trees+Mindstate and Process-Step-Analysis (detailed elsewhere) is the topic of managing externalized project data, or project-object-databases.

The Project Object Database:

Even in a situation where there isn't anything to track outside of the seemingly simple task you are asked to do, that task of tracking the object-relationships between parts of that problem->process->solution can be quite a juggling act, especially when you need to be able to externalize and check the parts and their relationships. For example, having a passive-reactive-reflective AI do a simple math word problem is (let's say) easily done. You put in the question, you get the answer. But let's add in externalization and confirmation; it is not enough to produce an answer in a black-box way (and this is a real concern for people using AI), show exactly what you did and how and why and how you confirmed it and confirmed that all the inputs and outputs to your deductive process are correct. In both cases the answer may be the same, but in the second case there is a vast book-keeping and signal-sending maze that must be gone through. And yes, each part of the maze in isolation can be done by the simple-reactive-reflective-AI component, but what architecture can do the entire process?

Though it is entirely possible that I am barking up the wrong tree with my dogged fixation on 'externalization,' an interesting part of project-tasks as separate from purely-abstract tasks is the formality of externalization. Let's go back to the very simple 1+1 task, forgetting about tea for the time being (can one ever really forget about tea...), and even forgetting about reporting and roles and all those other things we will have to account for later. Imagine your AI task really is as simple as  $1+1+1+1+1=5$ , but each '1' comes from a different source, in different data format, etc. In retrospect it looks very simple:  $1+1+1+1+1=5$ . But without context at the beginning of the process, if you didn't know it was going to be that simple, you would still need a larger process-step framework, architecture, which is most likely more involved than you would at first guess. What are the values? What are the operators? How many steps are there? Are the steps in the right order? How can you check? What thing from step one is supposed to go into step two? How do you check for errors? How do you find a mistake? What do you do when you find a mistake? What is the last step? (Process Step Analysis is a whole other important branch area of details.)

And the 'distance' from  $1+1+1+1+1=5$  to blue + red + orange + green + yellow = five colors, or darjeeling + assam + Earl Grey + Celan + Pu-erh = five cups of tea, or more steps involved, or a mix of numbers and words, and then adding in auditing and reporting and parts of the task modified and done by different participants: there needs to be an project-object, process-step-object, object-relationship 'database' in the architecture. No matter how good a given sub-task-process is at handling that task internally, projects on the whole require external, or externalized, data and processes.

### **Two-Brains and A Connector-Thing: Sharing Objects**

There are a number of marvelous oddities around the separation of skills, rolls, and tasks (and we need to be careful what we call these groups, or we may mistakenly lump them together or split them apart). For example, the models

that are not only the best but so far the only models that can handle-objects (I will try to summarize that shortly) are especially 'internal' in how they work; but the whole point of sharing things around a group-project-space where we work together on tasks is that those things have to be shareable. Just as people working together may have difficulty and take some time in explaining and documenting what they are thinking or seeing or realizing or remembering (or just notes from some event that only they attended). Just as you cannot readily share something somewhere deep in your brain, so a Large-Language-Model cannot readily share the details of how it is 'handling objects.'

The mammalian brain, perhaps by coincidence but perhaps not, also has a curiously similar structure.

So let us conclude with the sad but also somewhat comical image of the perhaps apocryphal problem that a person who has had their corpus callosum cut (for example to treat epilepsy) can have when getting dressed in the morning. Their right-brain (Using their left hand) does up the buttons on their shirt getting dressed for work...while at the same time their left-brain (using the right hand) undoes all the buttons...for whatever reason. To work together we need to share the details about what we are doing. And while we might like to imagine that tasks are done by "One Man, One Action!" thankfully we live in quite a different reality where tasks are done by networks and webs of participants and parts and processes, and while we like to argue about the semantics of labels and groups, whatever words we end up using (you call it 'the brain,' she calls it 'the collective of brains', they call it 'mind-soul-box,' whatever the agreement or disagreement in labels...) the objects must be shared between parts for things to work properly. It is not enough to 'train' one part or sub-part to 'just do' the task; the architecture of all the parts (however described and labeled) needs to 'learn' how to get that task done.

### **Never-Mind vs. Neurotic Worry-Monger**

A mirror, or a simple passive generative AI does not ask any of these questions. It has no project-state-mind-state. There can be light, no light, happy events, sad events; the mirror has no mind and is passively unconcerned. A normal computer function is likewise passively unconcerned; it is completely indifferent as to whether you get a truly delicious cup of tea or an error message or nothing at all. We want an AI that is very concerned, and scanning and tracking and testing and pondering all the parts of the project-space and their mind-space about what they are doing and how that impacts the other participants as well.

### **Zen-Mind, Feedback, & Mindful AI**

As a parting idea, sometimes no-mind is translated interchangeably with Zen-mind (and I have probably flippantly done the same myself). But there are some interesting subtleties that are likely relevant somewhere for AI. There

are, depending on the author, translation, school of thought, etc., various aspects of Zen-mind or mindfulness which seem appropriate to discuss for AI mindstate.

- awareness
- feedback
- compassionate & caring
- perspective
- not-distracted
- not-fixated
- not-distorted

To the extent possible, we want an AI-Architecture that is aware of feedback without being overwhelmed, that somehow manages to keep perspective and not be distracted, and that is concerned to not cause harm to others and to, if possible, be aware bad things happening to others as something that should not happen. And in my view the sectarian-zeal of fundamentalist relativism has shifted the overall discussion (which is a great discussion to have) too far in the nihilistic-hopeless-apathy direction: I would argue there is a lot of pragmatic low-hanging fruit where STEM and ethics and AI and some forms of compassion and empathy naturally overlap in realistic and practical ways.

Part of the occasional disagreement in semantics is around terms like 'unconcern.' There are some people who advocate for a kind of super-extreme anti-world, nihilistic, view of "mindfulness" where they argue for a kind of anti-meaning oblivion. I am not convinced that this is a coherent and meaningful approach, and perhaps if that is their goal they would not even disagree. More broadly 'unconcern' in a context of 'Zen Mind' (and the history of Hinduism and Buddhism) is not meant in a super-extreme way such as, if you are meditating and the person next to you catches fire then you should ignore their screams and empty your mind being completely unconcerned the health and possible suffering of everyone around you. Rather, the meaning of terms like 'unconcern,' or 'empty mind,' is usually in the context of having a balanced and healthy awareness and not being so extremely concerned with, or blindly fixated on, any particular tangent so that you can maintain a healthy macro focus on your state, the world's state, your fellow-participant's state, and see how to maintain and cultivate a harmonious and productive world.

### **System Collapse and Ethics**

As much mileage as we can get out of a practical model of system collapse will be extremely valuable, and there is more low-hanging fruit than might be suspected.

### **A Note on, and a Plea for, Making Useful Connections: STEM**

Please do not mistake my attempts to articulate the observation that some perspectives on, and applications of, math-logic are narrow (e.g. my sometimes disparaging terms "pure-math," "pure-land," and criticisms of "symbol

manipulation," etc.) as somehow meaning that math and logic either should not, or cannot be connected with other areas, or somehow that pure-research is in any way bad: it is not. When real life situations call for people with different backgrounds and areas of expertise to work together on a project, some people have prohibitively narrow ideologies and deliberately prevent the collaboration. But STEM itself is potentially more interconnected than people have usually assumed. The history of generalizing STEM and discovering or inventing connections between different disciplines and categories of systems across STEM (which has too many sub-areas to make a fully exhaustive acronym) is a fascinating topic that to some depth should be part of general education for all people.

As of 2023, we are teetering between paradigms with a real possibility that we will either fall back (into dogmatically believing that participants cannot collaborate) or drag on in a half-aware limbo. While it may be tempting to say that it was only very recently that people started thinking of a general STEM that connects all the different areas, saying this would be getting ahead of the starting gun. As of 2023 there is still no clear concept of a generalized STEM, and discussion of applications of and integrations of Math, Science, Data-Science, Engineering, Computer Science, Medicine, Epidemiology, Statistics, Logic, Linguistics, etc., bare all the hallmarks of a lack of familiarity with history and across disciplines. A crucial work in this area is the blessedly back-in-print Sir. Eric Ashby's "Technology & The Academics" which is a short and brilliant book everyone should read about the evolution of conceptions of science and science education in the late 1800's and early 1900's. Some of the most poignant history-shaping examples of this topic are laid out in the not-short biography of Alan Turing by Andrew Hodges, "Enigma" (which I can only hope everyone will read, as it is a phenomenal book on history, culture, technology, and WWII). Once you see these notions of STEM areas being connect-able or not-connectable, you should start to pick up on how other books will add to this story of intellectual history (such as the biography of Claud Shannon, "A Mind at Play," by Jimmy Soni, or "The Man from the Future," by Ananyo Bhattacharya about John von Neumann) or how a book may appallingly mangle and confuse events and concepts such as "The Idea Factory: Bell Labs and the Great Age of American Innovation" (not a recommended book). The culture, language, and psychology around STEM is fascinating and the very much still-in-progress development of understanding how parts of the world work is something we should collaborate and participate on carefully just as we should be working to make sure that physical planets can continue to be habitable for multicellular life.

### **In Closing**

We are accustomed to talking about a given very specific ai-model learning or training for a very narrow subtask, but we do not have a very good vocabulary for generalizing similar concepts so that we can zoom out to the bigger picture. The (perhaps) time tested routine of gathering data to train a model and using that model to do something is not a bad thing and may well be a general part of daily life for everyone in the future, a basic process just

like using a calculator, or a text-messaging program (...though I'm not sure if those will really be recognizable in the future. Up until recently the abacus and slide-rule were broadly used, but some tools do not last). We must not confine all imaginations to only thinking about ai and models as the single-reactive components they have been so far, used for the single reactive uses they have been used-for so far.

How do we teach an architecture to participate in a project and manage project-state and its own mind-state in a mindful way? How do we help an architecture to 'learn' to have the skills and abilities to manage a project-state and its own mind-state in a mindful way? How can we expand the set of jobs that AI can get done?

**Back to the main Object-Space paper, and ready for the next pass.**

We are hopefully closer to clearly discussing how in order for an AI to do a concrete task such a cut-up problem, that AI is going to need to use a variety of resources, or perhaps to be the intersection of a variety of resources, to weave together different kinds of solutions to different kinds of tasks and problems that are connected by the context of the overall project-task. We are hopefully closer to clearly discussing how doing project-tasks require architectures.

In all this we may now see an emerging question: "Ok, we need to, the AI needs to, keep track of stuff. What stuff? What kinds of parts are there in this space that need to fit together?" That question brings us, roundly roundly (henry-iv-part-1/Act1/Scene2/), back to the Howth or Alnwick Castle and Environs of Object Relationship Spaces, which is something of a wrapper for System and Definition Behavior Studies, which is something of a wrapper for better understanding the question of what STEM areas are and how they relate to each other and the rest of the world. What is science specifically, and how do we use it? Let's put the tanks on and dive back into the details. Here are some links. Hope to see you on another riverrun.

### **About The Series**

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

#### 4.3.31 Subroutine Stacking Measure: AI Learning

<https://medium.com/@GeoffreyGordonAshbrook/subroutine-stacking-measure-ai-learning-general-project-participation-bb342a2903c4>

Subroutine Stacking Measure: AI Learning & General Project Participation

2023.09.03

A Normalized measure of 'Intelligence' or 'Project Participation' and learning rate ability level in specific contexts.

A way to approach a standardized measure of learning related to problem solving, if not also an approach to problem solving itself, may be to employ a (perhaps normalized) measure of performance which may be useful as a proxy for:

- ability and capacity to learn
- Machine "Intelligence"
- General Project Participation (potential or ~kinetic)
- or even, assuming that capacity to learn is roughly equal, how conducive an environment is for learning across a spectrum of learning participants (including groups of participants)

#### Equation for Measure

Being empirically minded, I have much less interest in trying to anticipate-denovo what a given useful equation will be rather than suppose there is in principle (without any testing) a single-master-measure (likely? perhaps, but no data to go on yet).

We can start with two main parts to the equation, and a few design questions of course.

#### Two main parts:

1. a core-hub-kernel-seed something like this: create, save, and be able to re-use, a subroutine for part of a specific task/skill/ability set. A low-hanging-fruit context is STEM-math, which has a roughly well traced out branching tree of skill areas with various depths.
2. number-of-Increments [some operator]  
duration-of-learning-per-increment

**Design Question (and Phantom 3rd Part): Normalization, Halting, and Strange limits to learning**



What is a practical way to approach the end-limit for learning learning by a given system/participant/group in a given context (sorry if the descriptive term 'limit' collides with other maths jargon)?

There is hopefully a practical and tractable side to this as well as a ponderous side. Here is where a normalized measure may be helpful in cutting the gordian knot of knots of learning potential being theoretically infinite, and time being theoretically infinite, and the end of a calculation or process (the halting problem) being impossible to predict in principle. While those ponderous bits are there and interesting, in any real world situation there is a more tangible factor of 'this is taking longer than we can afford to wait,' and probably (that interesting if this is not the case) increasingly deep and difficult tasks may predictably take longer than shallower easier tasks.

A Mystery of 1:

Assuming we have such a normalized measure, zero means it can't learn anything, or you would wait forever to see if it ever got off square one (you could test this by timing how long your pet rock takes to learn to play A Suite for three Oranges (Prokofiev) on the bagpipes). Tuning a useful middle may take some empirical fiddling, and possibly (just as ~average standardized test performance mysteriously increases over time with H.sapiens humans) this will also need to be adjusted in time if too many participants become highly proficient. The mystery of zero is probably first to be seen here (a classic halting problem, etc.), but there may also be a mystery of 1. What happens when a system becomes so proficient as modular-task subroutine-making that it effectively never slows down in a reasonable amount of time? Aside from theoretical extremes, literally, how long should you wait to see if you can detect any slow-down before saying: ok, call it a 1 (a parallel problem to how long you should wait before saying: ok, call it zero.

AI vs. Other vs. Project-Context: Too Many Cooks in the Kitchen?

An interesting part of this measure is that it can just as easily apply to an AI or a H.sapiens-human (or any animal) or to groups of various size.

What is the optimal team size on a project to learn and operate well? Three might work better than one, but 3000k are unlikely to coordinate more quickly than 3. (Ever try to decide where to have lunch with 3000 people?)

Architecture-Learning vs. Model Training

How this might also overlap with the supposed-dichotomy of symbolic vs. subsymbolic I have not mapped out, but this at least also a topic where learning happens not only in a context of model-training, but also 'architectural learning' where the whole-ai-body matters as well as the model-training that goes into various 'parts of the brain.'

see *ai bodies and brains* paper:

<https://medium.com/@GeoffreyGordonAshbrook/ai-bodies-brains-solving-a-problem-4f0c59a2c3ca>

4.3.32 AI ALU Corpus Callosum

<https://medium.com/@GeoffreyGordonAshbrook/ai-corpus-callosum-rapunzels-corpus-callosum-1707b6a7cf02>

**AI ALU Corpus Callosum**  
Rapunzel's Corpus Callosum

~2023.08.16 g.g.ashbrook

## **Corpus Callosum for for AI Architecture**

**Context:** 'a union of opposites' vs. 'externalized distribution'

### **Strength of Bridging:**

- Symbolic / Sub-symbolic
- right brain / left brain
- system 1 / system 2
- All the flexibility of question-interpretation + and all the speed of ALU computation.
- 'good at identifying things'

A main theme or goal in the western tradition of alchemy was said to be the union of opposites.

The corpus callosum questions is infinitely open-ended in time in terms of what ways and trade-offs may be employed in specific instances and their generalizations for knitting together, pick your phrase, system-1 non-analytical and system-2 analytical processes, patterns, and aspects of a system, task, etc.

There is no single solution to speak of, but a hauntingly broad topic to make less-unintelligible.

### **Mo-tai-nai Invention: Philosophy, Practicality and Efficiency**

While an AI can write a script, or create layers of software, to solve a problem that the AI is not suited to generativity-guesstimate the solution to, it is 'not efficient' to over and over create new solution-architectures to the same sub-operations that it has and will routinely use, often taking time and experimentation to get each new approach (to the same, or essentially the same, problem) correct.

By analogy, it would be as if a cpu needed to create a new ALU by an evolutionary process literally every time it carried out a calculation, then threw it away and started tabula-rasa with the next arithmetic operation, or if a compiler did not have standard ways of handling the same routine ("subroutine") processes that it handles every time. Zen mind?

## **What is user-friendly for AI?**

How does AI like to approach problems?

### **The Recipe Question: Locations in a Process**

How does a given AI prefer to receive and accept parts of a process? While in some ways H.sapiens-human language trained LLM AI is just like H.sapiens-human, in other ways some tasks are easier, generating and debugging reGex expressions for example is easier for an AI, and counting a list of names is more difficult for an AI (in 2023). Tasks deeper in a Kasperov event horizon context are more difficult, but also 'tactical' single-layer tasks can be similarly difficult for AI and H.sapiens-human (e.g. starting from zero or 1 when getting items from some data-structure when lots of other details are buzzing around). In some cases a machine will prefer to exchange information in a reverse polish notation stack process, in other cases the machine prefers to handle natural language strings of the same serial expression and equations. Case by case, some answers may be surprising, and perhaps surprisingly interdigitated with the question of input structure (see the five trees) and the 'state' awareness of subroutine options.

Alice: "What size hex-wrench does this match in your toolbox?"

Bob: "I have a toolbox?"

The recent history focuses on two more technical areas

In neuro-sciences, system1 and system2 from Khanaman and Tversky, in computer science the perhaps incorrectly named, symbolic and subsymbolic processes. Neither of these may be 'correct' in completely matching this problem space, but the general area of problems space is likely at least a partial overlap.

The history of the topic becomes more murky, with early-science alchemy devoting much attention to the union and interplay of opposites in a more philosophical way than popular after the

enlightenment. and in the east an in/yo or yin/yang interplay of opposites,

Right and left brain was popular for a while, even the fascinating Julian Janes exploration of a Hamlet's Mill style mythology parsing, and the eternal "Drawing on the Right Side of the Brain" (I heard the first edition was best) but I am not sure of the current state of the science of this (right and left).

## **Externalization et al**

For example there are aspects elucidated by the context of Object-Relationship-Space based AI-architecture and OS studies which do bring in other contexts such as 'externalization' and multi-participant project elements, maybe including distributed, parallel concurrent and asynchronous aspects which are usually not in the above dichotomy contexts.

A guiding context here is more practicality and efficiency.

While an AI can write a script, or create layers of software, to solve a problem that the AI is not suited to generatively-guesstimate the solution to, it is 'not efficient' to over and over create new solution-architectures to the same sub-operations that it has and will routinely used (often taking time and experimentation to get each new approach (to essentially the same problem) correct).

By analogy, it would be as if a cpu needed to create a new ALU by an evolutionary process literally every time it carried out a calculation, then threw it away and started tabula rasa with the next arithmetic operation: zen mind. Or if a compiler did not have standard ways of handling the same routine ("subroutine") processes that it handles every time. Re-use of modular functionality is a core theme in computer science.

## **Process Memory, Algorithmic State**

This may also highlight another aspect or context of memory or state in systems, where a system can remember or not remember procedures and functions for operations.

## **Improvement**

A side topic here is finding better ways, processes, heuristic, algorithms for doing things. The history of sorting algorithms may be a mini-case study in this overall phenomena of improvement.

## Overall Process

1. the ability of the AI to not only write functions but to call on a library of existing functions/subroutines.

E.g. when handling different object/relationship types and or categories of types of systems (as in definition behavior studies).

Not just to ask it to use the tool, but to create a framework in which it can select an appropriate tool on its own based on what tasks it is set (see Mind-state for being persistently aware of, and share, available tools, provided or self-archived). This again can be done in either a 'symbolic' hard-coded 'GOFI' module or system, or a 'classify the task' sub-symbolic machine learning module could be used to try to accomplish this, as yet, speculative goal.

**Re-use of modular functionality is a core theme in computer science:**

- systematizing types of questions
- properties of Types of questions
- target: a set of questions that can be approached with an established set of tools and methods
- select a target space
- ways of error-checking
- not reinventing the wheel every time

## About The Series

*This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:*

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

4.3.33 The Vortex: Gender and AI's Scylla and Charybdis

<https://medium.com/@GeoffreyGordonAshbrook/the-vortex-gender-and-ais-scylla-and-charybdis-bcc155e26786>

The Vortex: Gender and AI's Scylla and Charybdis  
Group Monologue in an NLP Hall of Mirrors

Aside from a general plug for some excellent books by women...

**1. The Odyssey**

translated and prefaced by Emily Wilson (Author)

and wonderfully performed by Claire Danes

<https://www.amazon.com/The-Odyssey-Claire-Danes-audiobook/dp/B07GL4J6Z1/>

**2. Broadband by Claire L. Evans (2018)**

<https://www.amazon.com/Broad-Band-Untold-Story-Internet-ebook/dp/B074DGKWRG>

**3. Story of the ENIAC Six:**

Proving Ground: The Untold Story of the Six Women Who Programmed the World's First Modern Computer

by Kathy Kleiman (2022)

**4. Invisible Women:**

<https://www.amazon.com/stores/Caroline-Criado-Perez/author/B082ZRN17J>

(A hat-tip to Michael Wooldridge for recommending this.)

***What if participants in an imitation game ignored what was being said?***

As a case study in navigating obstacles and in the theme of multiple participants, during experiments with a general framework for (in the spirit of Prof. Skip Ellis's 2002 AI-Agent facilitated multi-participant meeting project, meeting with multiple AI agents (who had not been adeptly informed about the multi-participant framework for the meeting) took to creating, not responding to, all of the other participants.

Perhaps in yet another illustration of an "AI" (which is a slice of H.sapiens-human behavior) exhibiting an all too human behavior: we can see a 'group monologue' phenomena, which each man-splaining speaker in turn quite literally speaks for everyone else, creating a resulting transcript that without under the hood logging data is impossible to decipher in terms to which comments came from which speaker, as each AI speaker proceeded to write all other speaker's dialogue for them.

Perhaps sometimes the word 'your' should be emphasized when saying "Use your words."

**Between Scylla and Charybdis**

In these times peril, if we are not guided by the voices of wise women, we shall deservedly perish as the unsympathetic villains our destructive actions demonstrate us to be.

**Lines**

King Lear as Meeting Moderator:

"use my eyes"

"Come, let's away to prison. We two alone will sing like birds i' th' cage. When thou dost ask me ...So we'll live, ...and sing, and tell old tales, and laugh At gilded butterflies, and hear poor rogues Talk of court news; and we'll talk with them too- Who loses and who wins; who's in, who's out- And take upon 's the mystery of things, " Lear Act V, Scene 3 The British camp, near Dover.

The Council:

"When a wise man gives you better council, give me mine again. I would have none but naves follow it, since a fool gives it."

### **About The Series**

*This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:*

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

#### 4.3.34 Modeling Participant Architectural Learning in Five Trees Plus Mindstate

<https://medium.com/@GeoffreyGordonAshbrook/modeling-participant-architectural-learning-in-five-trees-plus-mind-state-36a98ee66655>

**Modeling Participant Architectural Learning in Five Trees Plus Mindstate**

2023.09.10-17 g.g.ashbrook

AI Architecture and Architectural Learning

### **AI Externalization and Project-Tasks**

Along with the themes of Architectural-Learning (vs. model-training) and ...

Let's look at some of the details of what kind of formal processes and sub-tasks are involved (and assumed) when asking or relying upon an AI or machine or automated or whatever participant to take up a task in a project.

Using the Woz-Office-Coffee-Bot, we could make a minimal task example that will nevertheless show how implicit a single project-task may be additional if hidden tasks or task-aspects, which nevertheless change the learning/skill/ability requirements for who/what can accomplish this task (and H.sapiens-humans are not oriented towards thinking of the world in terms of generalized-STEM project-tasks in a value-function-meaning non-system-collapse

set of interlocking contexts). For example, getting a list of preferred teas, finding out the prices and delivery dates, making a purchasing proposal, and reporting on the estimated cost. How this actual project task in externalized signal, steps, documentations, and actions that depend upon currently unknown information, differs from a single guesstimation of the overall outcome is crucial. Indeed, while there are cases where being able to guesstimate the overall outcomes is either necessary or advantageous, there are also cases where by virtue of trade-off being better at the project-management skills of the sub-parts is much more important. And where these two sets of skills and fitness profiles differ represents a very interesting fork in the road for where resources for investment and development should find baskets for their eggs.

### **Architectural Learning vs...**

Beyond stating and conceptually illustrating that 'a task' may appear to be the same as another but actually be quite another thing altogether, let us try to dip a bit into how one might model such a practical course of action.

### **Five Trees:**

1. In / Input / Context / Instructions
2. Out / Output / (Analysis type and details)
3. Content (subject-matter)
4. Process / Project-Tasks
5. Tests / Feedback

### **Plus:**

Mind-State: (or some kind of machine-state):

### **Notes:**

1. As with definition-behavior studies and object relationship space architecture studies, here there is no illusion of or allusion to a neat singular tidy solution.
2. Two sections, content and process, are possibly largely overlapping, but case by case hopefully a practical answer to questions-needed-to-be-answered can be found.
3. "Participant" vs. AI:  
While this may be an erroneous quagmire of semantics, the idea is to generalize task-doing regardless of the rank, status, id, personhood-ness, or whatever else, (sniffing dogs?), of who-or-what does the task. In this discussion, AI is the focus...but so is AI and H.sapiens-human team cooperation...and we still need to get around to describing H.sapiens-human learning for tasks which we have avoided doing for a few million years.
4. In: For a given task, assuming no input-tree or pipeline may be a large assumption, perhaps reminiscent of the ENIAC-paradox, where given a sufficiently structured input a problem could be done on the ENIAC, but that assumption fits squarely in the feasibility-of-method aspect of computer science: unstructured inputs are a stubbornly difficult problem, yet (if unable



to count) subsymbolic (including generative) AI is a powerful tool for this task. As a whole, when put into a full project context, in the larger balance of what is done by what kind of deep-ai or GOFAI-ai, deep-ai may be more intensively used on the input-tree side of the task, and in some output-tree reporting depending on what that is, when the content-tree and testing-tree locations are already known to be core-basal highly-conserved STEM areas such as known math calculations.

5. The term "Learning" here will be used for "architectural learning" as an abbreviation of a larger set of interrelated and largely synonymous (and to be honest largely not fully understood) terms: training, learning, skills, abilities, proficiency, performance, achievement, literacy, memory, mind, problem solving ability, cognition, calculation, benchmarks, baselines...etc. Listening out all the terms (assuming such a list of related terms is finite, which it may not be) (and where each term is no doubt controversial and hotly opposed with much adieu in its own right) repeatedly would make even a simple sentence very hard to follow. And depending on parts of speech and common phrases, other phrases may be used as well.

Even for the same task, model learning and architectural-learning, task process in one case and task process in a seemingly synonymous case, may substantially differ.

## **Area Details:**

### **1. In**

1.1 The ENIAC Paradox: Given enough structured input, even the original ENIAC might be able to do the computation, but given lingering ambiguities in input clarity, the same problem scope-creeps into requiring exponentially more resources.

1.2 Hofsteder's Gap

1.3 Context: Structured-ness of input, Instructions.

1.4 cli vs. fancier ui for the same task

### **2. Out**

2.1 Classification

2.2 Analytical Processes and Problem Solving

2.3 Subroutines and AI-ALUs

2.4 Process-Step-Analysis Reporting

2.5 Protocols and Requirements, Externalization

### **3. Content**

3.1 categories of Types of Systems

3.2 Object Relationship Spaces

### **4. Process**

- 4.1 Cutups
- 4.2 Herding Cats
- 4.3 Following Breadcrumbs

## **5. Tests**

- 5.1 A Tree of Tests
- 5.2 Unit Tests
- 5.3 Fallacy Criteria

## **Mind State:**

- 1. Memory
  - 2. AI Mind-state & Project Autonomy
    - Goals
    - Tasks
    - Tools
    - Questions
    - Overall Priorities and Values
  - 3. Time:
    - "What time of day is it?"
  - 4. Loops, Memory, Tasks, Resources
    - ~repl loop:
    - to remember that you can look things on line.
- short and long term layers:
- a two layer loop, usually shallow loop, periodically a longer loop...
- 2. Decentralized Mind-State: network-loops

## **Learning, Skills, Ability: Not a 6th Tree**

In a sense there is another related tree here, to understate the matter:  
Learning, skills, and abilities: What can your system do? However I recommend that we restrain ourselves from merely collecting a sack full of trees because we are tree-collectors, and pause to remember our goal: Modeling Learning

In some other contexts architectural learning may well be a tree, but in this context, we are defining learning in the context of a specific project task as the other five trees plus mindstate. Though this could be a strange exception, we probably as usual want to avoid adding 'tree of all possible architectural-learning forms' to its own definition: as amusing as positive feedback loop explosions can be.

## **Questions:**

- General System Collapse

- Comparing with a Non-Architecture Baseline
- primary task example: AI doing math problems
- secondary: ethics in projects
- participants: H.sapiens-humans, dogs, pigeons

**Also see:**

- AI ALU Corpus Callosum / Rupunzel's Corpus Callosum
- AI Bodies & Minds
- Modular Problem Spaces

**About The Series**

*This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:*

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

<https://medium.com/@GeoffreyGordonAshbrook/biology-psychology-math-ai-broad-or-ai-narrow-0e0a2a435ba8>

**4.3.35** Biology, Psychology, Math: AI Broad or AI Narrow

Biology, Psychology, Math: AI Broad or AI Narrow

**STEM integration, language, and non-monophyletic trees**

2023.10.06,15,21 G.G.Ashbrook

**Assignment One:**

Brainstorm: What are the top 5-10 items that you associate with Artificial Intelligence, and ask a friend or colleague what they think for the same topic.

**Planning AI: History, Amnesia, Priorities, & Distractions**

Reality vs. Attraction: AI, STEM & Pathology

Sciences Hard and Soft: AI & Distraction

List 1:

AI:

- Biology
- Math & Logic
- Psychology Etc.

Projects:

- Value

- Function
- Meaning
- Progress

vs.

List 2:

AI:

- horror
- shock
- popularity
- winning
- clicks
- wealth
- easy money
- bling
- fame
- violence
- entertainment

Themes in Collapse & Disinformation:

- amnesia
- bad choices
- failed processes
- failed communication

## **AI broad or AI Narrow**

Do we see AI (Artificial Intelligence, Machine Learning, Data Science) AI-ML-DS as a narrow offshoot of applied computer science, where chiefly one particular deep-learning technology is predominantly used by industry and 'hyped' by media?

Do we see AI as a broad intersection between many STEM (and perhaps non-STEM) areas that have grown together over time, including being 'non-mono-phyletic' meaning either it is one thing with multiple origins, or it is a collection of things with separate origins.

How are we using the terms 'broad' and 'narrow' in this context?

"Broad and Narrow" can be an especially important set of terms for how AI is described and considered before and after 2022 and the rise and spread of LLM GPT ANN deep learning models. From the status quo period from 2012-2022 people had gotten used to a 'normal' world where AI could only be conceived of as 'narrow.' 'Broad' or 'general' (not very well defined terms, more as arm-wavy

gestures) AI was considered either only possible in an alien distant future or impossible in principle.

But here broad and narrow are being used in a different context, not to describe the ability of a specific AI technology, but looking at the AIML field overall: do we take a narrow view ("It's just calculus. It has nothing to do with biology.") or a less narrow view of where AI has come from and whether it is interdisciplinary or not (whether or not we apply AI to a task or project that might be 'broad' or 'narrow' in other uses of those terms).

### **How Deeply Do you look? A chicken and Egg Problem of Terminology**

Why were Alan Turing, Claud Shannon, and John von Neumann, founders of not only computer science but AI specifically, "mathematicians" and not, well, computer scientists?

By analogy, if you are going to study tea, and what tea is, and connections to and from tea, what do you do when the first recorded mention of the use of the word "tea" is probably in the late 1600's (difficult to pin down for some reason), yet obviously tea (even if referring strictly to tea made only from the Camellia Sinensis leaf) is an international subject that extends through time and languages beyond 1600's England.

There is also a kind of chicken and egg problem, when trying to understand something for which the vocabulary we use now arose after the events you are looking at. In a number of STEM areas, and around the same time in the 1800's Gregor Mendal, George Boole, Charles Darwin, and Charles Lyell, were all developing ideas that we now describe using terms that did not exist when they were doing their work.

The same is maybe the case with trying to understand developments such as the 'invention' or 'discovery' of algebra (see: Ian Stewart's 'Significant Figures' <https://www.amazon.com/Significant-Figures-Lives-Great-Mathematicians/dp/0465096123> ) where different parts were added by different people in different places and times, making the whole story difficult to clearly describe using the same vocabulary across the whole story.

AI is particularly tricky for many of these compounding reasons and more, including the secrecy and confusion generated by WWII, where parts of the story are still considered 'national security' top secret, as well as cultural biases such as the gender biases making the early women programmers and Alan Turing himself outside of what could be discussed. E.g. Klára Dán von Neumann, John von Neumann's wife, was a great mathematician and directly involved in early computer and software development, yet is not mentioned at all in books attempting to focus entirely on the forgotten ENIAC programmer women...not to lay blame but to describe the deeply obscure and confused state of the history of computers. Klara dan von Neumann even wrote her own memoir, "A Grasshopper in Very Tall Grass" but to this day no one will publish it, even electronically. And these are the most famous contributors that we know that we

do not know enough about, let alone the many others who we know even less about.

As a last note on the difficult semantics of the history of AI, there is an overwhelming amount of indirect evidence all pointing to the idea that George Bernard Shaw's play "[Pygmalion](#)" was actually, overtly, a science fiction story about the future relationship between humans and the robot-AI servants they would create (along with parallels with social classes among people, gender, etc.).

This is so clear now, that Joseph Weizenbaum named the program 'ELIZA' after the 'living statue' mythical referenced character [Eliza Doolittle](#) in [George Bernard Shaw's Pygmalion](#), [a play] which was directly named after the greek myth about humans creating 'artificial' humans and then actually marrying and having children with them. (Also, Shaw's later play 'Back to Methuselah' even more overtly describes humans creating AI-robots, and uses the same term "**Pygmalion**" as the name of the artist-scientist who creates the 'artificial' people. And, according to Hodges, Shaw's 'Back to Methuselah' had a lasting impact on Alan M. Turing.)

But at the time the play 'Pygmalion' was written, none of these descriptive terms and concepts that we use now existed, and officially we still see "[Pygmalion](#)" or "My Fair Lady" as being 'just about people,' and you don't see it in the science fiction section of the book store. So it is an interesting example of one set of ideas caught between the lexicons of two different times (or across several times).

## Roots & Trunks

Even at the first official AI event in 1956, there were a number of areas that could have been, or could be, more parallel than convergent:

[https://en.wikipedia.org/wiki/Dartmouth\\_workshop](https://en.wikipedia.org/wiki/Dartmouth_workshop)

From the proposal: Machine Tasks they tried to figure out included:

- "use language,"
- "form abstractions and concepts,"
- "solve kinds of problems now reserved for humans,"
- "improve themselves"

And from wikipedia's summary: research topics discussed in the proposal included:

- computers
- natural language processing
- neural networks
- theory of computation
- abstraction and creativity

Note, 'computer vision' was not included.

### **Parallel vs. Overlapping:**

Both in 1955 when the proposal was made and in 2023, and so perhaps even more in 1955, it is unlikely that they were thinking of one single computer-robot-architecture and tech-stack of OS and software that would be used for all areas of study.

It may also be worth noting that to some extent, if not clear linguistically, some of the descriptive topics from 1955 have since been dropped from common discussion and descriptive language even if they have not been actually dropped from project tasks. And other 1950's goals may have been dropped from both (later descriptions and later goals).

"Form abstractions and concepts" are probably topics that people would be much more shy about talking about now, but this is partially semantics and cultural posturing. 'concepts' are the embedding-vector deep learning systems that have been successful, but this leads to much of the confusion around narrow deep learning vs. the still not well understood large language model success. Before 2023, and likely after as well, there was a concerted effort to not 'over promise or over-reach' to avoid major pull-backs from research and funding ("AI Winters" as they are called), and to focus on narrow concrete tasks such as image classification.

"Abstraction and creativity" may be an area where researchers and companies are not keen to advertise stimulating creativity (though after 2023 that may have changed compared with before 2022).

This 'walking on eggshells' sensitivity to how we speak about AI might be entangled with a narrowing of how we think about AI as well. There are probably many factors involved in changes over time from 1923 to 2023.

Another example of AI having more than one 'root' is where in the history of perceptions they describe the perceptron as originally not being thought of as a computer but as a machine. Because most computer-machines today are at their core a digital computer we may find it hard to shift our thinking back to when computers or digital computers were just one type of machine. This is also part of understanding the works of Norbert Wiener, and early science fiction.

The biography of Claud Shannon, *Mind At Play*, does a rather excellent job of showing what different technologies existed for doing different tasks before the creation of digital technology which ended up replacing many things.

### **Parallel Roots & Trunks in the non-monophyletic tree of AI**

Assignment 2: Pick areas to map out

Here is an (incomplete and imperfect, sorry) outline of different areas and disciplines that have contributed to, and converted into, AI.

Engineering & STEM:

- Calculating Machines

Statistics vs. Maths:

- Bayesian Machine Learning in 1718

linguistics and cryptography:

- codes
- types
- information-entropy

math:

- math and computer science
- Most of the founders were, by label, Mathematicians. In this context we have seen AI grow out of 'math.'

math and logic:

- Hilbert
- Russell & Whitehead
- Godel
- von Neumann
- Turing

perception:

- 1943 by Warren McCulloch and Walter Pitts

learning:

- statistical
- tree
- heuristic
- 'sub-symbolic'

psychology:

- James
- Freud
- Jung
- etc.
- cognitive psychology: Khanaman Tversky

neurobiology:

- perception
- memory
- plasticity



general biology:

- von Neuman
- genetics & DNA
- Both Turing and Von Neuman (and in a less clear way Shannon(e.g. genetics) made detours into Biology on their way to AI)

Economics:

- Game Theory

computer science:

- Church vs. Turing for digital
- 

Cybernetics vs. Digital Computer Science

- Norbert Wiener

Linguistics vs. Natural Language Processing

Math vs. Computer Science vs. Linguistics: Hilbert & STEM

Biology:

- Genetics & Genetic Algorithms: Learning
- The hypothetico-deductive method: learning, testing, perception

Codes, Information-Warfare, Disinformation, and Cybersecurity

Automata:

- John von Neumann
- John Conway
- Stephen Wolfram

Robotics:

- irobot
- [https://en.wikipedia.org/wiki/Rodney\\_Brooks](https://en.wikipedia.org/wiki/Rodney_Brooks)

### **Extended STEM**

AI, and bringing computer science more deeply into STEM, may help to generally extend STEM into areas important for society, life, and productivity. While we cannot now say what the future will require, we can speculate about how to fill in a more comprehensive set of STEM related areas for managing projects:

- [Definition Behavior Studies](#)
- Generalized STEM
- Intersecting Areas
- Project Context
- [Coordinated Decisions](#)
- [Object Relationship Spaces](#)

etc.

## Math and AI

The story of AI can begin with a math problem that we are still trying to solve. This is not a 'problem' in the sense of  $1+1=2$  but more a problem like trying to do algebra with roman numbers: Hilbert's Problems, though I may be overly-liberal in describing them as the 'problem' of math needing to be carefully formalized and integrated with (or at least to be able to be integrated with) STEM. We think of math, or like to think of math, or want to think of math, as rules and consistency and repeatability and clarity and translatability, but History is plagued by mathematicians and logicians being cryptic and vague and secretive and incomplete and falling much more into gang violence than, well, than math.

A wonderful portrait in miniature for the overall scenario, also with an AI connection, is one of Issac Azimov's iRobot humanity and AI stories about the 'telepathic robot' where one of the main characters is a mathematician who gets so lost in politics and intrigue and rank and career advancement and personal pride, and the authority-ness and rank-ness, of who says what and who listens to who, and worrying about his stupid white silk gloves, that he is never (by the end of the story, and the demise of the robot-ai in question) able to determine whether the AI is able to do a discrete math problem or not.

Numbers that Learn:

I have never seen a history of computation or AI that starts with bayes, or even mentions bayes.

Read: <https://www.amazon.com/Theory-That-Would-Not-Die/dp/0300188226>

## Two Vectors in AI Problem Space

A possibly useful and simple way to plan and discuss what project-task within and overall project or system or ai or ai-architecture is to look at two contexts, each of which can be looked at as a spectrum:

1. analytical vs. non-analytical
2. ascii vs. media

Even aside from project-task discussions and sticking only with 'traditional' single-function at tools, the different individual tasks may fall in interestingly different regions of this "chart" which might help to shed light on whether the right tool is being used for the right job.

More broadly this may be part of the debate over what exactly to include in AI, with some people wishing to use (or also wishing everyone else to use) the term in a very specific and narrower way (such as only for a subset of deep learning models), whereas other people use the term more broadly as in AI-ML including all known types of models associated with some kind of 'training' or fitting. Then again, 'models' does not seem to fit at all with many 'symbolic AI' systems (e.g. neither ELIAZA nor Mycin 'learned' or 'trained'), which for perhaps most of the history of AI were considered the only real approach to AI.

## Swap, Cut, AI

2023.10.19 Geoffrey Gordon Ashbrook

### 1. Swapping

"Easy things are hard." -> 'Problem 12= Flip 1 and 2 in "1+2=3" in this sentence.'

When this problem is made more general, such that any number (or characters in any character set) might be involved, and any sentence structure and presentation may be used.

e.g.

'Problem 12= Flip 1 and 2 in 1+2=3 in this sentence.'

'Problem 12: "Flip 1 and 2 in "1+2=3" in this sentence."'

'Problem 12: Flip 1 and 2 in one+two=three in this sentence.'

'Problem 34= Flip 3 and 4 in 3+4=7 in this sentence.'

'Problem 34= In this sentence, flip 3 and 4 in 3+4=7.'

'Flip 3 and 4 in 3+4=7 in this sentence, Problem 34.'

'In 3+4=7, flip 3 and 4 in this sentence as problem 34.'

etc.

Possible definition of concept-using 'AI' in computer science

Can we say that:

- For this class of problem, the task is not possible across a normal spectrum of cases and edge cases (e.g. only hard-wired or 'structured input' cases can be computed) without "structured input," such as:

1. specifying orders, delimiter and forms
2. hard coding a solution or process for a specific single case

### 2. Cutup

Let's take this one step further and make this a cut-up problem, where various (perhaps randomly divided) parts of this problem are given to different participants. What 'system' will be then able to do the task?

### 3. Define "Understanding"

This may be a case study in 'easy things are hard' and looking at 'natural' vs. 'artificial' processes in a world where that line has become substantially blurred compared with earlier times (e.g. a world without any automated math or known boolean systems etc.)

'Problem 12= Flip 1 and 2 in "1+2=3" in this sentence.' is very fast and easy for most people (possibly even birds and dogs), and LLM GPT AI can likely do this, but how about 'non-AI' solutions?

Can the context of object-relationships help here?

Is part of this that the item is a mix of data and instructions but not in a static format? What lexicon can help here?

How many such classes of problems, perhaps common in project-tasks, exist where there are perhaps definition-issues that make the problems intractable in certain 'analytical' modes. How does our vocabulary need to improve? Could this type of problem be a basis to STEM-define what can be meant by 'understand' such that the problem is not solve-able without "understanding"?

#### **An Empirical Comparison of Model Types: Understanding vs. System-1 vs. non-analytic vs. sub-symbolic**

Given a dataset (presumably in one language, though perhaps not), how would different types of models perform? Would any model aside from a huge LLM-GPT be able to do reasonably well? Can any reGeX-alone system do well?

Model types and system types:

In terms of trying to better define what properties this (or some better) problem may have that distinguishes it from more broadly defined types of problems that can be solved by a same-input-same-output function, how can we describe this?

The term 'understanding' may be forever too vague, perhaps reminiscent of 'hoping' for something about which we have no real information (a category that is a request for a defined category). Other existing concepts may be a better fit, perhaps Kahneman Tversky's system-1 system-2, analytic and non-analytic, but the range of machine-learning/ai models are not immediately categorized into two such categories. Even in the history of AI, a hand-crafted decision tree is GOFAI, yet an automated decision tree such as XGBoosted-random-forest type is...at least not described as GOFAI, but is a decision-tree not a decision tree? (are there sub-symbolic trees? or some other 'branches' of trees?

Disciplines:

Does this represent a case where linguistics, psychology, logic, math, computer science, engineering, etc., cross-over and cannot be excluded from an analysis of the problem?

(also see the swap case in the classic Kernighan and Ritchie The C Programming Language)

**Human, Natural, Knowable, Artificial, STEM, Integrated, Isolated**

If possible I would like to raise and introduce a context and topic here without getting lost in the labyrinth it presents us with.

Let's start with just three (you can image a ~triangle of sets if you like):

- Human Biology
- Knowability in STEM
- "Artificial" Technology

Two counterintuitive things have kept happening over time as H.sapiens-humans have tried out different tasks in a technological way to see if they are knowable(as in STEM) or unknowable( as in something not-STEM, perhaps something super-natural) (which could also mean not-yet-knowable (perhaps interestingly inverted from Karl Popper's not-yet-disproven)).

At the same time, especially visible in AI, we see the 'easy things are hard' phenomena.

This sometimes leads to a kind of 'boom and bust' or 'summer and winter' exuberance and pessimism around technology. People doubt that something can be modeled by a simple line, or as in the case with ENIAC, people doubted whether you could use vacuum tubes to do the same thing that other mechanisms did more slowly, and were flabbergasted when it worked! Why?

From 1943 (early computer and neural-network work) to say 1963, there was phenomena progress (which it might be important to put in a backdrop of a very substantial tech-driven change in the nature of daily life from ~1870-1970; see '[the rise and fall of American Growth](#)') and when the first generation of ('symbolic AI') system came out, people got lost in the contrasting vortex between 'sometimes easy things are STEM, sometimes easy things are hard.' and 'sometimes STEM can amplify other STEM, sometimes "generalization" is so elusive as to have its existence in doubt.'

There is a quasi-paradoxical set of ideas here, where H.sapiens humans are the measure of performance and success for AI-tasks, yet AI is (again, sometimes paradoxically) expected to do the tasks the people do in sometimes completely different ways...but sometimes not expected to be too different.

There are two ways in which we expect AI to be similar:

1. "Easy things are hard" We tend to expect that AI will have the same relative levels of difficulty with tasks as H.sapiens-humans.
2. That being good at one thing will 'naturally generalize' both for people and AI.

Deep Blue is a fascinating case, and something of an outlier to the systems that are usually covered by a 'what is AI' survey, given that it is so specific it is more or less not a general purpose computer at all, but rather a jumble of hardware and switches that occasionally works to do one very specific task.

Super-Polarized Natural vs. Artificial:

In some ways H.sapiens-humans take an extreme stance viewing some parts of the world as being the exclusive sphere of 'nature.' Here, if I can, I will sit on the fence and encourage dialogue between the both sides without 'joining a team.' Whether one invokes culture or mythology or aesthetics (none of which I view or mean here as pejorative ('bad') things), 'ear kissing' ideas (to use a Shakespeare phrase) such as "Science has no feelings! It's tough!" or "The soul has powers that math can never have!" are very attractive but tend to be polarizing and not helpful with exploring the fascinating intersection of generalized STEM in a strange world.

### **Artificial Networks vs. Natural Networks...or just "Networks"?**

If only there were ways of mapping the ebb and flow of what has sometimes been called the '[astonishing hypothesis](#)' (see below) that parts of life we think of as being too difficult to understand such as mind and many parts of biology are compatible with STEM. It is hard to predict when and where people are assuming this is either obvious or completely impossible. Perhaps like how what was once (before Friedrich Wöhler and Hermann Kolbe in the 1800's, as usual no one agrees on an absolute date) a chemistry-of-non-living-things and a separate chemistry-of-living-things simply became 'chemistry' and how the celestial physics beyond the moon and the terrestrial physics of earth (once defined as being impossibly different) simply became: physics (generally attributed to Newton's synthesis).

"The [astonishing hypothesis](#)" is a 1994 book by Francis Crick, the co-discoverer of DNA. And, did you know that Francis Crick is one of the PDP group members who co-authored the authoritative Neural Network Deep Learning book along with Geoffrey Hinton et al in 1986). In fact the section on neurons (not artificial vs. natural, just STEM neuron studies) in the foundational deep learning, artificial neural network, text was written by biologist Francis Crick, as a biologist, writing about neurons in general which includes biological neurons.

Yet, along with saying rude things about Ray Kurzweil, a standard requirement of proper books about AI is to make it very clear that deep learning Neural Networks and biological neural networks have nothing whatsoever to do with each-other. It is as though if any connection is allowed that this would somehow give permission for anyone to claim that AI-machines are full of human brains in jars. Over-reacting to sloppy science journalism runs the risk contradicting the historical record and making discussion of important and difficult topics even more difficult. (In my experience, it has made any conversation, or even raising the topic, literally impossible. The central dogma of AI seems to have become that there is an inviolable wall preventing even a discussion of both biology and AI. This is not consistent with STEM, or much of anything.)

### **Frontier: General STEM Integration**

An unexpected, if not counterintuitive, implication of AI being part of computer science (CS) and part of STEM, is that feedback from research into AI will, quite naturally, help us to learn more about various STEM areas. The

expected part may come if research into AI tells us about larger than expected gaps in our understanding of various areas.

Areas we do not yet know how to frame:

- Statistics
- Linguistics
- 'mind' and consciousness
- dynamical, nonlinear, chaotic & stochastic, fractal, etc.
- game theory
- automata
- general ecology
- networks
- math-notation
- "symbol manipulation"
- psychology

AI not only re-raises questions about the relationship between 'hard' sciences and 'soft' sciences, but also different areas within STEM that are not simply the same thing. Areas of STEM work together, but that does mean that math simple is logic which is physics which is statistics which is chemistry etc.

Another possible area, if not openly discussed, is that before the later rise of AI the status quo consensus was that so-called 'soft' sciences would just go away, fading away into nothing. Psychology had become pharmacology. Ecology had become a marketing term for 'organic' foods (an ironic shadow-term from before 'organic' compounds were generally accepted to even be made of normal physical matter). Sociology and communication studies were given up as post-modern 'arts and crafts' utter nonsense, less practical than golf course design as an academic subject. But with the rise of the internet, social networks, mass disinformation campaigns, problems with hate speech, trolling, and bewilderingly successful recruitment of the general public into extremist cults (religious, political, or just strange), and then the rise of LLM GPT AI in 2023, and with the re-discovery of the "cognitive-psychology" and "behavioral-economics" of Khanaman and Tversky, it was not so obvious that these subjects could and should be all be ignored and written off.

Linguistics, still not on most people's radar, is probably ripe for a renaissance. What happens when people react to advances in Natural Language Processing, aka AI, by asking: So what about the current science of language? What does science tell us about language and the mind? Who wants to say? "A science of language and the mind? You have to be joking. No one's even tried to look at either of those scientifically since the 1800's. Important people have been ignoring that question for a long time, and now so should you!"

And yet who has the naive-bravery to openly say (perhaps reminiscent of some old story about an emperor and his wonderful new clothing) that we have no idea what language is. We don't know what mind is. We don't don't know what consciousness is. We'd basically decided in the late 1900's that mind, consciousness, culture, and even meaning, don't exist at all; they are just

illusions for fools and marketing campaigns. And we don't know what statistics is. And there is no conception yet of a generalized STEM. And whole parts of STEM are completely ignored, such as Computer Science (as relates to Math and Logic (Null values?)) and project management.

Note: Project-Management and Agile are still not broadly seen as relating to STEM (as with computer science itself) so there are still more changes to come as future integrations build on, and ratchet up, the forms and functions of AI.

## **Science Fiction**

Both for fiction and for non-fiction we would do well to read what people wrote before 1970. While it is amazing if not puzzling that the way people lived between 1870 and 1970 changed so dramatically, it is perhaps more puzzling than amazing that intellectually and culturally the world after 1970 the world contracted in such a reactionary way.

In the books written before 1970, many of them out of print and not online, we have an amazing window in the lost continent of pre-1970 thinking. One area of this is early science fiction.

Issac Azimov's stories about "robots" (AI) are a wonderful discussion of social and design issues that anyone involved with AI (so...everyone?) should read.

According to the biographer of John von Neuman, in a nice section on self-reproducing machines and automata, the author mentions that a co-inventor of basic concepts for machine self reproduction was actually Philip K. Dick.

A good starting point may be the more than twenty volumes of collected stories that Azimov edited and published (not his own, but ones he recommended) by various authors, too many of whom are now not remembered:

- Theodore Sturgeon
  - James Blish
  - Fredrick Brown
  - Frederik Pohl
  - A. E. Van Vogt
  - Alfred Bester
  - Poul Anderson
- and many, many, more.

## **History and Timelines**

Assignment 3:



Make a timeline of items you think are important.

Example:

#### **Historical Roots**

- PDP vol.1 vol.2
- Turing & Biology
- Von Neuman & Biology
- Rosenblatt & Biology
- System1, system2 & 'cognitive psychology'

#### **Pre-"AI"**

1956-1971: First Explicit Era of AI

- 1966: ELIZA
- 1968: SHRDLU, "blocks world"
- The First Internet ~1961-1969

1971-2011: AI Underground

- "Perception"
- PDP, Hinton, Lecune
- Hofstadter, Mitchell
- 1976 [https://en.wikipedia.org/wiki/Computer\\_Power\\_and\\_Human\\_Reason](https://en.wikipedia.org/wiki/Computer_Power_and_Human_Reason)
- 1980: Searl & The Chinese Room

2012-2022: Biology Reborn: PDP Hinton's

- Narrow Deep Learning

2023: Object-Handling General AI

- OpenAI: "They're not laughing anymore."
- Open Models

Pre-AI and Shrouds of Secrecy

- WWII national secrets
- Western blindspot for women
- Western blindspot for software

**Looking at who was alive when and how that shaped the narrative.**

### Death-Day Parties

A N Whitehead: Died 30 December 1947 (aged 86)  
A M Turing: Died 7 June 1954 (aged 41)  
John von Neumann: Died February 8, 1957 (aged 53)  
RonaldFisher: Died 29 July 1962 (aged 72)  
Klara von Neumann:Died November 10, 1963 (aged 52)  
Norbert\_Wiener: Died March 18, 1964 (aged 69)  
Walter Pitts: Died 14 May 1969 (aged 46)  
Warren McCulloch Died 24 September 1969 (aged 70)  
Bertrand Russell: Died 2 February 1970 (aged 97)  
Frank\_Rosenblatt: Died July 11, 1971 (aged 43)

#### [1971]

John Mauchly Died January 8, 1980 (aged 72)  
Vannevar Bush: Died June 28, 1974 (aged 84)  
Kurt Godel: Died January 14, 1978 (aged 71)  
Philip K. Dick Died March 2, 1982 (aged 53)  
Isaac Asimov Died April 6, 1992 (aged 72)  
Karl Popper: Died 17 September 1994 (aged 92)  
Alonzo Church Died August 11, 1995 (aged 92)  
Konrad Zuse: Died 18 December 1995 (aged 85)  
J. Presper Eckert Died June 3, 1995 (aged 76)  
Claude Shannon: Died February 24, 2001 (aged 84)  
David Rumelhart Died March 13, 2011 (aged 68)

#### [2012]

#### [2023]

Martin Davis Died January 1, 2023 (aged 94)

### Can't Join the party..because they are still alive in 2023:

- Ken Thompson
- Brian Kerinigan
- Terry Winograd

### What should your brain do when you are reading an old physical book that other people suggest should not exist?

In attempting to find (cheap) old copies of history books, I obtained a copy of PDP Vol. 2 by Mclelland, Rummelhard et al.

A. Paging through this book really made it sink in, for me, that biology and psychology ARE part of the history of AI. The book IS biology and psychology AND it is the oft-cited foundational work in deep learning.

This is not 1940's early speculation, this comes after the rise and all of symbolic AI, during an AI winter (which is perhaps why it isn't calling itself AI).

B. I ordered this (as a used old book) because it was referenced as foundational in newer books, but I'm not sure how I ordered it, because it doesn't appear to exist in web-searches as a past or present book.

Volume 1 appears to be in print,

<https://www.amazon.com/Parallel-Distributed-Processing-Vol-Foundations/dp/026268053X> but there is no clear reference on Amazon to vol 2. There are some vague offers for "both" volumes, but the complete title of volume 2 does not appear anywhere in Amazon.

If you do a google search on the title in firefox...

<https://www.google.com/search?client=firefox-b-1-d&q=mit+%22Parallel+Distributed+Processing%2C+Volume+2%3A+Psychological+and+Biological+Models%22#ip=1>

the results do not suggest that this is even a book.

If you do the same search in chrome:

<https://www.google.com/search?q=mit+%22Parallel+Distributed+Processing%2C+Volume+2%3A+Psychological+and+Biological+Models>

you get the same results, BUT you also may see a sales-add bar on the side for the book, on Amazon!

[https://www.amazon.com/Parallel-Distributed-Processing-Vol-Psychological/dp/0262631105/ref=asc\\_df\\_0262631105/](https://www.amazon.com/Parallel-Distributed-Processing-Vol-Psychological/dp/0262631105/ref=asc_df_0262631105/)

So Amazon does sell this book!!...but it does not come up any ANY searches.

e.g. if I click on the author, [James L. McClelland](#), the page lists only vol 1 and his other books.

Why is finding history so difficult? How many other books are lurking hidden in the semantics of searches?

### **Perceptrons, Biology and AI**

Along with perfunctory attacks on Ray Kurzweil, which seem to be an obligatory social ritual for people writing books on AI, another common dance routine is saying that artificial neural networks, and perceptrons, have nothing at all whatsoever to do with biology. Zero. Nothing. Well, maybe they were inspired by biology in the 1940's but nothing more. Nope. Not a single thing. No connection. No overlap. Absolutely no.

I had heard and read this so many times that it became ingrained in my head, as well as my being dramatically shamed by very credentialed co-workers for asking about comparative network behavior studies comparing biological network dynamics, such as neuroplasticity, and other networks including artificial neural networks.

Please take a minute and read the wiki on Frank Rosenblatt, even just a skim.

[https://en.wikipedia.org/wiki/Frank\\_Rosenblatt](https://en.wikipedia.org/wiki/Frank_Rosenblatt)

Just like the PDP books, this work is not just distantly influenced by biology, it is biology. To quote one sentence "In 1970 he became field representative for the Graduate Field of Neurobiology and Behavior, and in 1971 he shared the acting chairmanship of the Section of Neurobiology and Behavior." He was the leader in mainstream neurobiology and behavior studies. Not 'artificial robot behavior,' biology.

But something has happened (perhaps starting around 1971) that changed many fields of research so deeply that it is difficult to claw our way back to history.

Everyone I have met and read today is completely convinced that AI has zero connection at all whatsoever to biology, or psychology, or neurology, etc. So what is the best way to juxtapose that with this written history?

**The problematic state of AI and computer science history:**

1. many untold stories
2. many stories told only many years after the fact
3. widespread sloppiness and apathy
  - out of print books
  - books of which there is no clear record they ever existed
  - factual inaccuracies:
    - naming errors (sounds like bad code!)
    - wildly erratic dating of events (without sources) mixing up events before, during, and after WWII. (including in books otherwise well researched such as

A Mind at Play: How Claude Shannon Invented the Information Age Paperback - Unabridged, July 17, 2018 by Jimmy Soni

**- Disinformation:**

Doubtless when Sara Turing and Kathy Kleinman were researching their books, they met lots of disinformation and discouragement (some troubling accounts are included in in Kleinman's excellent book).

**Many thanks to those who recorded the history for us to have!**

To Joseph Weisenbaum's role in helping the story of the ENAIC-6 women programmers to be tracked down and told!

Without his help, Kathy Kleiman might not have had the resources or support to play her role in both researching the story and organizing events and recognition while the ENIAC six were still alive.

Proving Ground: The Untold Story of the Six Women Who Programmed the World's First Modern Computer Hardcover - July 26, 2022  
by Kathy Kleiman (Author)

To Alan Turing's Mom who put enormous effort into writing his first biography.

To Andrew Hodges for putting together a fabulous history, and to Durmot Turing for continuing to dig into the story while in ~living memory.

Integrated Stem:

Hard-Science, Soft-Science, Not-Science

? Math-logic:

Hard Science:

- physics

Soft Science

- "creativity"

unknown"

- language

- statistics...

- categories of types of systems

### **Thorny Questions: Astronomy, Astrology, Divination, Prediction & Quantizing Measures of Change**

In Durmot Turing's book on the history of AI there is a small picture of a Sumerian clay tablet used for making astronomical and astrological predictions.

This is probably off the far end of what can be clearly included in the discussion, but how do systems and languages and codes for historical, often not entirely STEM, systems fit into history?

For example, while discoveries of 'the oldest' are bound to keep changing, some of the oldest written symbols include the binary system of the i-ching in the east (said to have impressed Leibnitz), which was part of an older technology for measure change (much as Leibnitz contributed to what became an ultimately binary digital calculus based system for computing change).

How do millenia old technologies such as the i-ching, or astronomy charts, or the antikythera device, relate to types of questions that we ask machine-learning systems today in pursuit of technologically assisted answers?

Did people ask ELIZA the same questions they asked Tarot Decks, Runes, Geomancy frameworks, dream-interpreters, tea-leafs, horoscopes, and that people now ask chatbots?

How much do we know about what alchemists were studying in analyzing the 'union of opposites'?

For example people still (somehow) speak of digital computers as 'manipulating symbols' as though performing divination.

Another way to approach this set of questions might be, how much has feedback from using our tools taught us (*H.sapiens-sapiens*) about ourselves? Can we see more clearly how we use language and what we mean by 'symbol' than 5, 10, 15 thousand years ago?

Do we view language the same way in 2023 as we did in 23 or -2023?

Do we view STEM the same way as we did? (There the answer should be a more clear change over time...even if we cannot see the future trajectory.)

### **Cognition, Learning, and Non-Monophyletic Trees**

As a final twist, let's look at how hominids do not like thinking about non-monophyletic trees. A classic example from biology, or botany in particular, of non-mono-phyletics is the lilly. The term 'Lilly' refers to a number of very different plants that happen to have similar looking flowers. In terms of genetic evolution these plants are not related to each-other. So in such cases, 'science' says that 'the lilly' does not really exist, because it does not have one root, it is not monophyletic (one historical tree). So it is a kind of mistake, a cultural illusion.

In the case of biological structure such as eyes or wings or limbs, this is less simplistic, as it is not so easy to say that 'eyes' or 'leafs' do not exist at all because they evolved in parallel from multiple different origins. And yet many people are inclined to say just that.

Or another branch of this science-language-psychology is perhaps illustrated by the fascinating 'a tomato is a fruit' phenomena, where in a kind of dogmatic literalism people try to make a cargo-cult collage of cultural language terms against jargon from randomly selected disciplines. And tragically people will often conclude that anything not covered by a jargon term of measurement (not intended for general language use in all situations) does not exist at all.

And so with STEM and with AI, we have a convergent or inverted tree structure where different areas grow together into a natural form. The H.sapiens-human mind very much does not like non-monophyletic trees, and will go to great lengths to avoid dealing with them, so what will AI minds (some of which are images of H.sapiens-human population language minds) do with similar concepts? How will learning from learning about learning learn about what is difficult to learn?

Assignment 4:

Research Something, what do you find?

Example:

### **No Volume of Founder's Letters**

At least for John Adams and Thomas Jefferson (or as Princeton's computer science department may have called them, John Anderson and Thornton Jennings) we have a volume of letters for them to express their thoughts. And even with these and more letters and biographical materials there are still massive gaps and unknowns in our understanding of who these people were and what they thought and did.

In the case of the history of computer science, it is almost as if it is a practical joke: what if the most obscure people, under national secrecy, had perhaps everything or nothing to do with some of history's most important, transformative, and integrating, technologies?

#### **The Founders:**

- Kurt Godel (logic, math, and computer science)
- Alonzo Church (Lambda Calculus, functional programming)
- John von Neumann (misc computer science, hardware, software, math, etc.)
- Klara dan von Neumann (programming and computer science)
- Claud Shannon (boolean logic, communication theory, AI theory, robotics)
- Betty Shannon (math & misc)
- Alan Matheson Turing (a bit of everything)

Depending on who you ask, on what date, any of these people had either everything or nothing to do with anything or everything with complete or zero communication with anyone or everyone (about everything or nothing). Despite that there are, in 2023, probably, still people alive who have direct (or slightly indirect) knowledge about what happened between these people, we literally have no idea what happened between the founders.

Maybe Claud Shannon was deeply involved with AI and even named the field, maybe he had nothing to do with it. Maybe Claud Shannon and Alan Turing were friends who worked together on information-theory (given that Alan Turing invented and used information entropy in the UK before Shannon wrote his synthesis) or maybe

it was a total coincidence and they never discussed it or anything of any technical nature. When they had tea every day at bell labs, maybe they just talked about the wall-paper, or stared at each-other in creepy silence, we have no reliable information.

Maybe the von Neumanns knew and worked with Kurt Godel (they were both at princeton IAS) or maybe not. The official 'witness' story is that Kurt Godel would mysteriously regularly simply walk into John and Klara von Neumann's house, sit and read a book silently for a few hours, and then leave. (Again, this is a classic perception-riddle: If there is a witness, then obviously that means someone other than von Neumann or Godel was also there. Godel was famously shy and paranoid, and von Neumann was working on top secret projects, so obviously nothing could be expected to be said while a third party witness happened to be there. Did Godel ever say anything when no-one was there to hear? We don't know because no one would have been there to hear. AH!)

Maybe von Neuman helped Alan Turing get to Princeton to do his doctoral because he noticed what Turing was studying, or maybe it was a random fluke: he accidentally helped Turing but otherwise didn't know Turing existed. This might sound ridiculous, but this is the line of investigation laid out in great detail by the fabulous Andrew Hodges (in his biography 'Alan Turing The Enigma'). Hodges is thankfully frank about what he actually has evidence for vs. what might seem reasonable to assume (unlike other biographies and histories where weaving a compelling narrative is the game). As with computer science in general: nothing is safe to assume.

Maybe Turing, while at the nexus of Princeton, bell labs, and the intelligence-cryptography center, working with Godel, Shannon, Church, and von Neuman, or maybe they had no conversations about anything ever.

Maybe John von Neuman met and consulted with Alan Turing while 'creating' the 'von Neumann Architecture' computer design that Alan Turing was already implementing in the UK, or maybe they never even met after the war, and it was a total coincidence. Maybe they both turned to studying biology together in the 50's, or maybe it was another complete coincidence.

Perhaps the only written evidence of any interaction between these figures (after the 30's) around computer science is also one of the strangest and most difficult to interpret. On his deathbed, in the hospital, with the brilliant Klara von Neumann answering his letters, finishing his writings, and managing his affairs while he was dying and incoherent: note, this is late 1956 or early 1957, John von Neumann gets a letter from Kurt Godel (a person notoriously paranoid and non-communicative, not known to be working with von Neumann or Turing or on anything computer-science related) asking for von Neuman to share his thoughts on Alan Turing's research and NP-completeness in computer science. Time-line alert: np-completeness is not supposed to have been discovered until 1971 (with no attribution to Godel, or Turing, or von Neumann). <https://en.wikipedia.org/wiki/NP-completeness> How are we supposed to interpret this letter? Are we supposed to believe that the completely



non-communicative, "paranoid" and loony, Kurt Godel, just randomly wrote a letter to someone (who was dying) he didn't really know, about someone else neither of them knew, about work that didn't exist yet, in a field he wasn't working in? That sounds like an unlikely hypothesis. But we also don't have any information to go on to hypothesize anything else. We don't even know if he was really writing to John, he could have been writing to Klara, who (if they had been working together) he would have known to be an excellent person to ask about computer science and math, that was her speciality.

Note: The story of this letter comes from the Biography of John von Neumann  
The Man from the Future: The Visionary Life of John von Neumann  
by Ananyo Bhattacharya (Author)

<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/1324003995>

### **Rebutting the automatic war-is-cause thesis**

A common causal assertion is that computers, the internet, AI (or any given part of STEM) came "from the military" or "out of the war" or "because of the war effort."

Nuance aside, this tangled case of the founders of CS and AI may be a counter-example to the common broad-brush argument that WWII 'caused' computers to arise (with the presumption that more "war" would "cause" more "progress"). WWII started after the founders did their foundational work and started collaborating. The war is a key factor in not only the lack of continued communication between the founders but also, if indirectly and not conclusively, their premature deaths (and in the cases of perhaps Godel, Shannon and Church, making them too terrified to ever talk about anything with anyone).

While we cannot say what would have happened if there were no WWII, the walls of secrecy and isolation that separated the founders is clearly not something that helped progress in CS and AI.

Am I saying that von Neumann and Turing should not have helped in the war effort, essentially sacrificing their own lives for public service? No, not at all; read Timothy Snider's "Black Earth"

<https://www.amazon.com/Black-Earth-Timothy-Snyder-audiobook/dp/B014X6Q80M> if you have any doubts about the clear need to mobilize against the Axis forces). I am trying to focus on the specific question of whether war created collaboration between the CS-AI founders leading to progress. In this case, as we have almost improbably no evidence of any collaboration and so no claim that there was a community of founders, the answer appears to be a very obvious: no. There is also a lot of incorrect recounting of computer developments that actually happened after WWII and or not part of WWII projects (such as ENCIAC related projects) that when incorrectly stated make it sound like all computer history happened during and as part of WWII, but that is not what happened.

For example, Vannevar Bush, sometimes portrayed as a military man who then shaped computing on behalf of the military, built the first useful [differential analyser](#) with [Harold Locke Hazen](#) at [MIT](#) between 1928 and 1931 (not during WWII,

not connected with the Military). Most people involved in the history of computer science have a long life of involvement and interest completely outside wars and the military.

Counter-examples might be people like Grace Hopper, whose career with computer software was directly shaped by her posts during WWII.

WWII did happen, and the pair of WWI and WWII shaped and reshaped the world probably more deeply than we can comprehend. But it is incorrect to characterize computer science, the internet, and AI, as simply appearing, like Athena from the head of Zeus, out of war, military, and violence as though otherwise there would have been no cause for anyone to imagine them.

#### Appendix 1: Data Dump

Andrew W. Appel, the "Eugene Higgins Professor and Chairman of the Department of Computer Science at Princeton University" (according to the back cover) thankfully edited and published a facsimile of Turing's 1938 thesis, along with a few commentary essays and miscellaneous remarks. This is a bit of an appendix of (more) Enigmas to the previous Enigmas.

There is a mostly parallel collection of materials online here, all about Princeton's connections to Turing, a few other figures, and ENIAC, <https://www.princeton.edu/turing//alan/> . A routine, affirmative, self advertising, anniversary celebration; a great opportunity to post archival material. Everyone wins.

Returning to the book, this is one of the strangest history books I have ever seen.

The suggestion is that Veblen, Turing, Church, Godel, and von Neumann all worked together to build computers in the US after WWII, in some kind of US-UK computer project (that excluded Bell Labs) and brings Konrad Zuse into the story add mind-bending levels of confusion and suggestion.

As far as any normal history:

1. There was, sadly, no connection between the US and UK post-war computer building project. Saying that Turing 'Did not contribute much' to the US computer building movement is an understatement: There was no US-UK computer movement; and Turing wasn't involved at all in the US movement that anyone has yet suggested. So cattily saying that 'other people were more involved than Turing' is simply bazaar. Everyone knows Turing had, tragically, zero involvement.
2. There was, sadly, no UK/US knowledge and recognition until after the 1940's of Zuse's work. He was not (that I know of) part of Operation Paperclip or brought into NASA or Bell or IAS.
3. I have never even heard it postulated as being possible that the people we retrospectively call 'the founders' actually were working together to build computers in the US and UK.

Is the author talking about building the computer industry? Is the author talking about academic pure research? What is the author talking about? If, confusing language aside, the author/editor is talking about soft-ware oriented pure research and not a computer-machine building industry, then A: why does he mixed hard-ware builders into his discussion, and B: if there are so many people who laid out the foundations of computer science so much more so than Alan M. Turing...who are they? What did they say and do, and where is this alternate history?

Turing is mostly known for his 'On Computable Numbers' paper of 1936 where he essentially invented computers (the Turing Machine) as part of addressing a Hilbert problem about the nature of math and logic, but this is his 1938 thesis on overcoming "Godelian incompleteness."

While the princeton website is a routine archive, the book edited by Appel is unclear in basic intent. The book is entirely a somewhat random dump of files about Turing, but the introduction by the author is both sour and contradictory. Appel writes "But as significant as Turing is for the foundation of computer science, he is not the only scholar whose work led to the birth of this field." But this is a book about Turing, and also not his computer thesis. So what is this book about?

And what does Appel mean by "this field"? By lumping together non-digital, non-electronic, non-program-running, non-software machines together with boolean digital electronic software running machines, it throws the whole context into complete confusion.

The editor also quotes Andrew Hodges, but in an apparently obscure attempt to contradict Hodges without actually providing any details that contradict Hodges. Hodges wrote that Turing (who had famously terrible handwriting and messy notes) hired a professional typist to type up his thesis after getting feedback from Alonzo Church about what to do in a next draft. Appel says: No. Hodges is completely wrong, people often left spaces in documents to write formulas by hand. That non sequitur bit of trivia may be true, but it does not contradict Hodges claim that Turing hired a typist, or got feedback from Church. What exactly is Hodges supposed to have gotten wrong, according to Appel?

The main mystery-item however is this line: "The great engineers who built the first computers are well known: Konrad Zuse (Z3, Berlin, 1941); Tommy Flowers (Colossus, Bletchley Park, 1943); Howard Aiken (Mark I, Harvard, 1944); Prosper Eckert and John Mauchley (ENIAC, University of Pennsylvania, 1946)."

Enigma: How is it that a book written by the head of computer science at Princeton, with no other purpose than to associate Princeton with (or celebrate Princeton's role with) ENIAC and Alan Turing, not only dismisses and denigrates Turing (and his biographer) but mis-spells the names of BOTH ENIAC creators: J. **Presper** Eckert, "**Pres**," and John **Mauchly**. Every history book I have found

repeats the same passage, over and over: "J. Presper Eckert. Pres, everyone just called him **Pres. Pres** Eckert."

John Adam **Presper** Eckert Jr., or J. **Presper** Eckert, or **Pres** Eckert. "Prosper?"

"Prosper" is a rare alternate name for a Shakespeare character: Prospero in The Tempest, as in:

The Tempest - Act 3, scene 3

ALONSO O, it is monstrous, monstrous!  
Methought the billows spoke and told me of it;  
The winds did sing it to me, and the thunder,  
That deep and dreadful organ pipe, pronounced  
The name of Prosper. It did bass my trespass.

<https://www.folger.edu/explore/shakespeares-works/the-tempest/read/3/3/>

How does the head of computer science at Princeton not know the names of the creators of the ENIAC, and the EDVAC, and the BINAC and the first US Computer Company: UNIVAC (and champions of the Women programmers who gave software a solid start).

So, it is a bit unclear what is going on with this book. Why did they publish telegrams to Turing and his return postcard, and a paper about Logic, Alanzo Church, and Kurt Godel, along with mis-spelled and inaccurate names of the ENIAC creators and a random line about there being more important people than Turing in the history of the field of computing?

Yet another twist, to make perhaps an Enigma pretzel: The beginning meets the end, with some odd connection in the middle.

#### 1. Type Theory:

According to Andrew Hodges, the story of John von Neuman and Alan Turing begins with von Neumann's interest in the Theory of Types.

#### 2. Transfinite Type Theory:

According to <https://www.sciencedirect.com/topics/mathematics/turing-thesis> :

" Turing considered several natural ways in which ordinal logics could be constructed: (i)  $\bigwedge P$ , obtained by successively adjoining statements directly overcoming Gödel incompleteness at each stage  $\alpha$ ; (ii)  $\bigwedge H$ , a form of transfinite type theory; and (iii)  $\bigwedge G$  (after Gentzen), obtained by adjoining principles of transfinite induction and recursion up to  $\alpha$  at each level  $\bigwedge G(\alpha)$ ."

And, according to the wiki on Turing's logic/types paper:

"Martin Davis states that although Turing's use of a computing oracle is not a major focus of the dissertation, it has proven to be highly influential in theoretical computer science, e.g. in the **polynomial time** hierarchy.[4]"

[https://en.wikipedia.org/wiki/Systems\\_of\\_Logic\\_Based\\_on\\_Ordinals](https://en.wikipedia.org/wiki/Systems_of_Logic_Based_on_Ordinals)

Remember, what did Godel write to John and or Klara von Neumann about? He asked about Non-polynomial time completeness and Turing's research.

(Note: Martin Davis sadly passed away in 2023.)

3. While it is more a tangle of questions than answers, we are starting to have at least the formation of an historian's question: While it is catty and disorganized this very odd book represents a nexus of connection between nearly all the parts that previously we had no way to connect. And while it is not clear, the 'claim' of this book by Appel is that at Princeton in the 1930's there was an active hub of computer science planning and formulation involving Turing, Church, Godel, von Neumann, Veblin, Lefshetz, Newman (and 11,000 other numbered but not named people) with insinuated connections to the rest of the computer science world.

If more from Martin Davis than Appel, we have one more puzzle piece in the link between Turing, Godel, NP-Completeness, and the von Neumanns. We can hypothesize a kind of shadow computer science consortium including the known founders and others with a specific thread of types, NP-completeness, lambda calculus (Church) and Godel's system.

Note: The ever elusive Claud Shannon was also at Princeton IAS (the institute for advanced study), and given that Appel went so far as to hyperbolically associate with > 11,000 (un-named) experts, it is curious why he left out Claud Shannon.

#### Appendix 2:

For another interdisciplinary survey through AI I very highly recommend

"Possible Minds"

<https://www.amazon.com/Possible-Minds-Twenty-Five-Ways-Looking/dp/0525557997>

#### Appendix 3: An Apple a Day

##### The Founders:

- Kurt Godel (logic, math, and computer science)
- Alonzo Church (Lambda Calculus, functional programming)
- John von Neumann (misc computer science, hardware, software, math, etc.)
- Klara dan von Neumann (programming and computer science)
- Claud Shannon (boolean logic, communication theory, AI theory, robotics)
- Betty Shannon (math & misc)
- Alan Matheson Turing (a bit of everything)

While it is a popular sport to take potshots at Godel for being "needlessly" afraid that something might happen to him, let's look at the fates of our founders.

Turing: Died under mysterious circumstances and inexplicably without official investigation. Now known to have been involved in UK, US, military intelligence after the war. Now widely considered to have been assassinated.

John von Neumann: Died possibly of cancer in his 50's, under 24 hour military guard in complete isolation because of his intelligence and military status. (Note: While von Neumann's biography puts a positive spin on this care and attention from the US government and military, Shannon's biographer mentions this military isolation as something that would have terrified Claud Shannon and which, unlike the gadfly von Neumann, Shannon would have shied away from and avoided at all cost.

Klara dan von Neumann: Official Cause of death: suicide by drowning.

Kurt Godel: Died horribly of starvation (while working at Princeton IAS) fearing assassination by poisoning.

Alonzo Church and Claud Shannon lived to die of natural causes in old age, living 'normal' lives.

Claud Shannon, who was eventually pulled by von Neumann into intelligence work, was reclusive and afraid to say anything to anyone about anything he did.

Alonzo Church lived until 1995, telling the world nothing about the amazing story he had a front seat to. Why is that?

Considering the momentous topic, amazingly little has been written about the founders. Turing, though now it seems popular to 'backlash' against any attention given him, is no longer completely obscure. But it is extremely difficult to get information about the founders, and even harder to guess their collaboration (with wildly contradictory anecdotal accounts, e.g. whether von Neuman traveled to meet with Turing in the UK or not). In the afterword to the rather long Hodges biography he is open about how only indirect scraps of information were available, attributing this to mathematicians mostly being obscure figures and no effort at the time put into preparing records for future interest they never predicted.

Appendix 4:

*Operations Research vs. Agile & CS*  
*Computer Science & Operations Research*

*Another lead in the ironically cryptic "Alan Turing's Systems of Logic" book, edited and introduced by Andrew W. Appel, is Appel's reference to 'the new fields of computer science and operations research' as being, rather unclearly, either prefounded or post-founded by either one group of named people, another group of*

unnamed people, or perhaps yet another group of un-named people (which does make you wonder what exactly he is trying to say, and why he isn't just saying it). Aside from the bafflement of his chronology, it is fascinating to bring "Operations Research" into the narrative.

As three dates for some timeframe, the first two from the UK's Operations Research Society <https://www.youtube.com/@Theorsocietypage>

1916 OK Organizing shipping Defense, creating procedures and methods  
1938 "Operations Research" coined  
1948 Rand Corp. founded in U.S.A. (after WWII)

At least at first glance, and I may be completely wrong of course, it appears that there is a kind of parallel between the relationship between Math and Statistics as there is between CS & Operations Research.

Statistics vs. Math

While most people probably associate statistics with math related to probability, the term (if only a historical note) literally comes from the term 'State' as in government or administration. Statistics is, of course, math, but more specifically it is math used for management and administration, or governments, or States, hence: statistics.

Operations Research likewise appears to be a very STEM related mix of administration and management, and government, and military-defense, specific methods and procedures (not really clear what to call it). In the US, where the term "Operations Research" is profoundly uncommon (especially in books about computer science or computer science history), as in the UK, the roots appear to be military-defense focused, and at least after WWII in the US the Rand Corporation is considered an example of Operations Research field world (which again has a distinctly military-planning-forecasting theme).

In 2023 Operations Research apparently still exists but perhaps with a less military specific theme, just as 'statistics' is hardly only used by governments.

Operations Research (OR) is a fascinating part of the overall STEM puzzle, in part because it appears to make connections between STEM and project and production management (whereas STEM is only maybe embracing Agile literally more than a century after early OR focused on project management), and because OR has somehow has not connected back to either computer science or STEM.

*There also appears to be possibly a connection to 'systems' thinking in OR, which for whatever reason has been slow to become acceptable in non-new-age circles (generally STEM and 'systems' in the US are considered incompatible).*

*And there appears to be the classic WWII & Postwar military-secrecy theme. As with other question-marks in the history of computer science, Rand seems to be a stereotypical von Neumann mystery.*  
[https://www.rand.org/pubs/research\\_memoranda/RM1019.html](https://www.rand.org/pubs/research_memoranda/RM1019.html)

So maybe, through von Neumann there was a close early computer science community involving:

- Princeton IAS CS & and Operations Research
- Rand and Operations Research
- Alan Turing in the UK
- Los Alamos (?)
- MIT
- Carnegie Mellon
- Penn
- etc. (not meaning to exclude anyone here)

But we do not know, aside from very scattered clues and very unclear suggestions (or just non sequitur editorial clutter not meant to suggest anything) from Princeton.

## **Appendix 5: Hinton on the Biology-AI Connection Question**

In this speaking event at the University of Toronto, Geoffrey Hinton (and Image-Net creator FeiFei li!) speak.

<https://www.youtube.com/watch?v=E14IsFbAbpI>

1hr:48min:12sec

Geoffrey Hinton and Fei-Fei Li in conversation

10,151 views Premiered Oct 7, 2023

This week marked the inaugural session of the Radical AI Founders Masterclass, featuring a dialogue between AI luminaries, Geoffrey Hinton and Fei-Fei Li. Held at the MaRS Discovery District auditorium in Toronto, the conversation was hosted by Jordan Jacobs, Managing Partner and Co-Founder of Radical Ventures, to delve into the profound ethical landscapes, societal shifts, and the transformative potential of AI.

Around ~20 minutes in there is a section where Geoffrey Hinton very explicitly describes his approach and work as "building a bridge" between biology and abstract technologies, making technologies that work closer to biology. And that his background is in psychology, the mind, and the brain.



It could of course be claimed that Hinton is wrong and does not know what he is talking about, but his consistent remarks about the facts of his well published background cannot be denied; in the professional disciplines and perspectives of Geoffrey Hinton, his deep learning technologies come out of, and from an integration with and closeness to, biology.

Appendix 6:

See:

<https://www.sciencedirect.com/topics/mathematics/turing-thesis>  
<https://www.folger.edu/explore/shakespeares-works/the-tempest/read/3/3/>  
[https://en.wikipedia.org/wiki/John\\_Mauchly](https://en.wikipedia.org/wiki/John_Mauchly)  
[https://en.wikipedia.org/wiki/J.\\_Presper\\_Eckert](https://en.wikipedia.org/wiki/J._Presper_Eckert)  
[https://en.wikipedia.org/wiki/Martin\\_Davis\\_\(mathematician\)](https://en.wikipedia.org/wiki/Martin_Davis_(mathematician))  
<https://en.wikipedia.org/wiki/NP-completeness>  
<https://www.amazon.com/Deep-Thinking-audiobook/dp/B06XWLY5XS/>  
<https://web.archive.org/web/20120901152639/http://www.math.ucla.edu/~hbe/church.pdf>  
<https://www.amazon.com/Possible-Minds-Twenty-Five-Ways-Looking/dp/0525557997>  
[https://www.amazon.com/Parallel-Distributed-Processing-Vol-Psychological/dp/0262631105/ref=asc\\_df\\_0262631105/](https://www.amazon.com/Parallel-Distributed-Processing-Vol-Psychological/dp/0262631105/ref=asc_df_0262631105/)  
<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/1324003995>  
[https://en.wikipedia.org/wiki/David\\_Rumelhart](https://en.wikipedia.org/wiki/David_Rumelhart)  
[https://en.wikipedia.org/wiki/Pygmalion\\_\(mythology\)](https://en.wikipedia.org/wiki/Pygmalion_(mythology))  
[https://en.wikipedia.org/wiki/Pygmalion\\_\(play\)](https://en.wikipedia.org/wiki/Pygmalion_(play))  
<https://www.amazon.com/Mind-Play-Shannon-Invented-Information/dp/147676669X/>  
<https://www.amazon.com/Proving-Ground-Untold-Programmed-Computer/dp/1538718286>  
[https://en.wikipedia.org/wiki/J.\\_Presper\\_Eckert](https://en.wikipedia.org/wiki/J._Presper_Eckert)  
<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/B09M2LTKSH>  
<https://www.amazon.com/Parallel-Distributed-Processing-Vol-Foundations/dp/026268053X/>  
[https://en.wikipedia.org/wiki/Rodney\\_Brooks](https://en.wikipedia.org/wiki/Rodney_Brooks)  
<https://www.amazon.com/Learning-Internal-Representations-Error-Propagation/dp/B00CC2EWC6/>  
[https://en.wikipedia.org/wiki/Differential\\_analyser](https://en.wikipedia.org/wiki/Differential_analyser)  
<https://www.amazon.com/Black-Earth-Timothy-Snyder-audiobook/dp/B014X6Q80M>  
[https://en.wikipedia.org/wiki/J.\\_Presper\\_Eckert](https://en.wikipedia.org/wiki/J._Presper_Eckert)  
[https://en.wikipedia.org/wiki/Isaac\\_Asimov](https://en.wikipedia.org/wiki/Isaac_Asimov)  
<https://www.amazon.com/Rise-Fall-American-Growth-Princeton-ebook/dp/B071W7JCKW>  
[https://en.wikipedia.org/wiki/Philip\\_K.\\_Dick](https://en.wikipedia.org/wiki/Philip_K._Dick)  
[https://en.wikipedia.org/wiki/Alonzo\\_Church](https://en.wikipedia.org/wiki/Alonzo_Church)  
[https://en.wikipedia.org/wiki/Terry\\_Winograd](https://en.wikipedia.org/wiki/Terry_Winograd)  
[http://ghn.ieee.org/Oral-History:Claude\\_E.\\_Shannon#National\\_Research\\_Fellowship\\_at\\_Princeton.3B\\_switching\\_publications](http://ghn.ieee.org/Oral-History:Claude_E._Shannon#National_Research_Fellowship_at_Princeton.3B_switching_publications)  
[https://en.wikipedia.org/wiki/Joseph\\_Weizenbaum](https://en.wikipedia.org/wiki/Joseph_Weizenbaum)  
[https://en.wikipedia.org/wiki/Kurt\\_G%C3%B6del](https://en.wikipedia.org/wiki/Kurt_G%C3%B6del)  
[https://en.wikipedia.org/wiki/Vannevar\\_Bush](https://en.wikipedia.org/wiki/Vannevar_Bush)

[https://en.wikipedia.org/wiki/Rob\\_Pike](https://en.wikipedia.org/wiki/Rob_Pike)  
[https://en.wikipedia.org/wiki/Ballistic\\_Research\\_Laboratory](https://en.wikipedia.org/wiki/Ballistic_Research_Laboratory)  
[https://en.wikipedia.org/wiki/Robert\\_Griesemer](https://en.wikipedia.org/wiki/Robert_Griesemer)  
[https://en.wikipedia.org/wiki/Ken\\_Thompson](https://en.wikipedia.org/wiki/Ken_Thompson)  
[https://en.wikipedia.org/wiki/Back\\_to\\_Methuselah](https://en.wikipedia.org/wiki/Back_to_Methuselah)  
[https://en.wikipedia.org/wiki/Dartmouth\\_workshop](https://en.wikipedia.org/wiki/Dartmouth_workshop)  
<https://www.audible.com/pd/A-Grasshopper-in-Very-Tall-Grass-Trailer-Podcast/B09VT8HV34>  
<https://www.amazon.com/Significant-Figures-Lives-Great-Mathematicians/dp/0465096123>  
<https://www.youtube.com/@Theorsocietypage>  
<https://www.youtube.com/watch?v=8bAKJufDWso>  
<https://www.youtube.com/watch?v=ILWbaWrjgU4>  
<https://www.amazon.com/Linear-and-Nonlinear-Programming-International-Series-in-Operations-Research--Management-Science-228/dp/3030854493/>  
[https://www.rand.org/pubs/research\\_memoranda/RM1019.html](https://www.rand.org/pubs/research_memoranda/RM1019.html)  
[https://en.wikipedia.org/wiki/Kl%C3%A4ra\\_D%C3%A4n\\_von\\_Neumann](https://en.wikipedia.org/wiki/Kl%C3%A4ra_D%C3%A4n_von_Neumann)  
[https://en.wikipedia.org/wiki/Charles\\_Lyell](https://en.wikipedia.org/wiki/Charles_Lyell)  
[https://en.wikipedia.org/wiki/Konrad\\_Zuse](https://en.wikipedia.org/wiki/Konrad_Zuse)  
[https://en.wikipedia.org/wiki/George\\_Boole](https://en.wikipedia.org/wiki/George_Boole)  
[https://en.wikipedia.org/wiki/Artificial\\_neuron](https://en.wikipedia.org/wiki/Artificial_neuron)  
[https://en.wikipedia.org/wiki/Charles\\_Darwin](https://en.wikipedia.org/wiki/Charles_Darwin)  
[https://en.wikipedia.org/wiki/Walter\\_Pitts](https://en.wikipedia.org/wiki/Walter_Pitts)  
[https://en.wikipedia.org/wiki/Alfred\\_North\\_Whitehead](https://en.wikipedia.org/wiki/Alfred_North_Whitehead)  
[https://en.wikipedia.org/wiki/John\\_von\\_Neumann](https://en.wikipedia.org/wiki/John_von_Neumann)  
[https://en.wikipedia.org/wiki/Dennis\\_Ritchie](https://en.wikipedia.org/wiki/Dennis_Ritchie)  
[https://en.wikipedia.org/wiki/Ken\\_Thompson](https://en.wikipedia.org/wiki/Ken_Thompson)  
[https://en.wikipedia.org/wiki/Grace\\_Hopper](https://en.wikipedia.org/wiki/Grace_Hopper)  
[https://en.wikipedia.org/wiki/Ronald\\_Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher)  
[https://en.wikipedia.org/wiki/Claude\\_Shannon](https://en.wikipedia.org/wiki/Claude_Shannon)  
<https://en.wikipedia.org/wiki/ELIZA>  
[https://en.wikipedia.org/wiki/Frank\\_Rosenblatt](https://en.wikipedia.org/wiki/Frank_Rosenblatt)  
[https://en.wikipedia.org/wiki/Konrad\\_Zuse](https://en.wikipedia.org/wiki/Konrad_Zuse)  
<https://www.youtube.com/@Theorsocietypage>  
<https://en.wikipedia.org/wiki/SHRDLU>  
[https://en.wikipedia.org/wiki/Norbert\\_Wiener](https://en.wikipedia.org/wiki/Norbert_Wiener)  
[https://en.wikipedia.org/wiki/Warren\\_Sturgis\\_McCulloch](https://en.wikipedia.org/wiki/Warren_Sturgis_McCulloch)  
[https://en.wikipedia.org/wiki/Walter\\_Pitts](https://en.wikipedia.org/wiki/Walter_Pitts)  
[https://en.wikipedia.org/wiki/Warren\\_Sturgis\\_McCulloch](https://en.wikipedia.org/wiki/Warren_Sturgis_McCulloch)  
[https://en.wikipedia.org/wiki/Computer\\_Power\\_and\\_Human\\_Reason](https://en.wikipedia.org/wiki/Computer_Power_and_Human_Reason)  
[https://en.wikipedia.org/wiki/John\\_McCarthy\\_\(computer\\_scientist\)](https://en.wikipedia.org/wiki/John_McCarthy_(computer_scientist))  
[https://en.wikipedia.org/wiki/Gerald\\_Jay\\_Sussman](https://en.wikipedia.org/wiki/Gerald_Jay_Sussman)  
[https://en.wikipedia.org/wiki/Frank\\_Rosenblatt](https://en.wikipedia.org/wiki/Frank_Rosenblatt)  
[https://en.wikipedia.org/wiki/Patrick\\_Winston](https://en.wikipedia.org/wiki/Patrick_Winston)  
[https://www.chessprogramming.org/Edward\\_Fredkin](https://www.chessprogramming.org/Edward_Fredkin)  
<https://www.chessprogramming.org/File:ShannonMcCarthyFredkinWeizenbaum.jpg>  
<https://www.amazon.com/Broad-Band-Untold-Story-Internet/dp/0593329449/>  
<https://www.amazon.com/Theory-That-Would-Not-Die/dp/0300188226>

<https://www.amazon.com/Pioneer-Programmer-Jennings-Computer-Changed/dp/1612480861/>  
[https://github.com/lineality/Online\\_Voting\\_Using\\_One\\_Time\\_Pads](https://github.com/lineality/Online_Voting_Using_One_Time_Pads)  
[https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)  
[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)  
<https://en.wikipedia.org/wiki/Mycin>  
[https://en.wikipedia.org/wiki/W%C3%B6hler\\_synthesis](https://en.wikipedia.org/wiki/W%C3%B6hler_synthesis)

**Authors you should read:**

Douglass Hofstadter  
Raymond Kurzweil  
Dermot Turing  
Melanie Metchel  
Michael Wooldridge  
R.G. Mulgan  
Andrew Hadfield  
Herman H. Goldstine  
Joel Shurkin  
Daniel Kahneman  
McClelland, Rumelhart, and the PDP Research Group  
Sara Turing  
Andrew Hodges  
B. Jack Copeland  
Sinclair McKay  
John Ashbery  
Sir Eric Ashby  
Brian W. Kernighan  
Dennis M. Ritchie  
Ian Stewart  
Robert J. Gordon  
Kathy Kleiman  
Claire L. Evans  
John Brockman  
Jean Jennings Bartik  
Sharon Bertsch McGrayne  
Gordon Welchman  
Hobson Lane  
Ian Goodfellow  
Francois Chollet  
Shakespeare  
Gary Kasparov

**How broad or narrow is AI, or the field of computer science?**

If AI and computer science more broadly are part of a larger synthesis of pure and applied STEM fields, one that will redefine and refocus fields such as linguistics, statistics, biology, psychology, sociology, and even the modeling and the scientific method itself, AI may be broader than we can imagine. And

while it is periodically predicted that everything that will be discovered has been discovered, we may be lounging at the foot of transformations that significantly reshape daily life, just as [life was transformed between 1870 and 1970](#) (see 'the rise and fall of American Growth')

### **About The Series**

*This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:*

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

<https://medium.com/@GeoffreyGordonAshbrook/jellyfish-spider-crab-ai-modular-architectural-learning-9b7492d68394>

#### **4.3.36 Modular Architectural Learning: Jellyfish Spider-Crab AI**

Jellyfish Spider-Crab AI: Modular Architectural Learning

2023.10.17-28 g.g.ashbrook

##### **Intro:**

My sister's four year old son picked out a birthday, or April fool's day, present: a fidget spinner with six jointed arms. Right off, I imagined it as a robot with completely modular nodes, off doing tasks, and I wondered how it might work. A robot jellyfish spider-crab!

"Mr. Spidercrab, please get my pencil."

"Ms. Jellyfish, could you see if the mail is in?"

"Mr. Spidercrab, go wake up Gladis and tell her the tea is ready."

"Ms. Jellyfish, please set a five minute timer and tap me on the shoulder when it's done."

I also imagined these 'modules' (each bending leg joint) as being, in future, able to reconfigure. For example, instead of one big spider, reconfigure into two smaller spiders. Or have two smaller spiders merge into one. This adds an element of ad-hoc organization (where there is no fixed set of participating modules) to the already decentered scenario.

You can have a minimal module that only moves, or you can add as many other sensors and abilities as...size and resources will allow.

Why 'Jellyfish'?

Jellyfish, or cnidaria <https://en.wikipedia.org/wiki/Cnidaria>, have (depending on squabbles over exact terminology, jargon, and semantics) no "central nervous system" in quite the same way that some other organisms do have clear central nervous systems. A jellyfish is closer to a colony of separate single-celled organisms that coordinate. So this is a nice segue into thinking about how individual modules can coordinate without a single hardwired centralized structure.

Why "Spider-crab"?

The choice is a bit arbitrary, but a spider-crab fits the anatomy rather well, one small ~round body in the middle surrounded by long segmented leg-modules: basically a bunch of legs that meet somewhere in the middle. (And I am keen on Rust for programming which has crustians as their mascot...and the gift came with a great big red crab craft-art card which I opened first...adorable crabs everywhere.) So crab it is!

### **Minimal First Experiments & Thought-Experiments:**

Barring another dark age, the components to build (or have fun trying to build) such a modular robot keep getting better, more resource efficient, smaller, safer, and cheaper (why do operating systems and browsers keep moving in the opposite direction?). So building a modular-AI-robot may, now or soon, be feasible as a school project.

### **Minimal example:**

- We can use a 3-legged beastly with just one module/cell/node per leg/arm: Basically three modules in a triangle. There are a gazillion ways the modules could 'move like an elbow bends' relative to each-other (or relative to the 'hub' in the middle, if there is a separate 'hub' at all. For our purposes here it does not matter exactly which design is used. (Coming up with and analyzing the advantages and disadvantages of different configurations is a great project-set and reminiscent of biology.)

### **Test 1: Stand up, please.**

- The solution/goal is likely to move all legs "down," in any sequence (no particular sequence) probably.

### **Test 2: Move towards a light source. ("Mothra")**

### **Test 3: Move away from a wall.**

- The jellyfish spider-crab is against a wall, perhaps the 'floor' is on an incline towards the wall.  
- The job, the task, the test, is to move away from the wall.  
- The solution may be to move the two wall-facing modules at the same time. This will require some but minimal planning and coordination.  
(there may be a more minimal goal to come before this...)  
(Simply moving all legs at the same time may work too.)

### **Give Each Robot Small-Instructions vs. Give The Robot-Team Whole-Tasks**

While one or more people can probably figure out a plan for what the robot should do, and hard-code a robot that can do that (not a bad project in and of itself) the idea-set here includes having the modular-AI team-of-modules get a higher-level task instruction and to figure out 'on its own' how to do that, with the added challenge that this requires all the modules to work together, to coordinate decisions, in order to accomplish the task.

### **Blunt Powers & Finer Distinctions**

An interesting timeline demarcation here is ~2022. If in 2020, or 2012, or 2002, you picked the goal of doing such an experiment with a starting instruction for the AI such as. "It's rather dark in here, I think. And I think there's a light-pull over by the sofa, or maybe across from the window. It sure would be good to have more light in here." It think it is safe to say that in pre-2022-times most people would find this instruction and goal for a robot (let alone a modular-AI conglomerate) to be completely beyond what AI could reasonably do, even though flipping a light switch (or pulling it) is significantly simpler than the Woz Coffee Test.

<https://www.youtube.com/watch?v=MowergwOR5Y> (posted Mar 2, 2010)

But after Open AI's gpt LLM models, we now have AI that can (however imperfectly and inconsistently) conceptually and spatially and temporally understand and model even fuzzy natural-language input and make viable plans for how to carry out a task.

(Unless this text gets somehow sent back in time) 'current' time being after 2022, each of our AI-modules can have all the sophistication of an LLM large language model as well as an array of sensor data and mechanical-operations (movement, light, sound, etc). Each node can easily make a feasible plan and (if less trivial) interpret the sensory data in a spatial-object project-task context. This is a leap forward that nearly every book written before 2023 (and perhaps most books published in 2023, as it takes so many years to write a publish a book) that I have seen flatly state this range of concept handling and language using abilities as impossible either for the foreseeable future or impossible forever in principle: yet since 2023 they are a practical reality available with any internet connection. This is a big leap in blunt AI power, but there are still important fine details.

Our modular-task is a (perhaps 'cut up' ish) distributed multi-step task that passive-reflective-reactive AI models have no means to do directly and simply all by themselves. This is an architectural-learning task. What architecture and abilities do the colony of modules need to be able to accomplish the task? What project-state/mind-state is needed? What 'body' is needed? What project-object-database is needed? What externalization is needed? What AI-Architecture/AI-OS is needed?

For 'future' people after this is being written, these questions may sound rather dull. But at a time when ~half of this task is shovel-ready, and half the task is...so close there must be a way, this is a delightful brink-of-exploration topic.

### **Coordinated Decisions, Generalized**

While "AI-modules communicating about a task and agreeing on a plan" can be described and paraphrased in many ways, I will sometimes use "coordinated decisions," as a phrase that is hopefully clearly descriptive and not opaque jargon.

While it may seem rather narrow and not a big deal for three nodes to coordinate standing up, this is a kind of foot-in-the-door or even a slippery-slope into a potentially bigger area. This does not necessarily mean that a given robot-task ('pick up a pencil,' or 'walk around') is going to be more difficult (though some goals may unexpectedly be so), but there are potential real-world applications in many areas that branch off from the ability to effectively coordinate decisions that are across a range of topics and professional-fields not always thought of as being connected.

The topic of coordinated decisions is big, fascinating, and amazingly largely unmapped. It does not yet appear to have been recognized by H.Sapiens-humans as a general area of STEM, rather different parts of it are dealt with in isolated areas usually lacking systematic study. Terms such as voting, elections, project planning, leadership, negotiation, mediation, project management, cell signaling, etc., are considered separate, mostly as non-science quagmires, as opposed to branches of the tree of STEM.

This is a very work-in-progress github repo of mine that started just looking at the question of whether it is feasible in principle to use one-time-pads in a vote-over-a-network election, but has expanded to explore coordinated decision systems more generally. If nothing else, hopefully this (huge mess) serves as a list of leads and potential topics for what areas are involved in General Coordinated Decisions:

[https://github.com/lineality/Online\\_Voting\\_Using\\_One\\_Time\\_Pads](https://github.com/lineality/Online_Voting_Using_One_Time_Pads).

Just as digital computers are built on zeros and ones but turn out to be much more versatile than just saying 'zero' or 'one' in a given instant and location, decisions and signals may seem elusively overly simplistic: a signal travels from here to there. But the consequences of a problem-space or project-space of signals, like the consequences of zero and one, or the consequences of four-valent-carbon (organic chemistry), are not narrow.

As usual I would also like to gesture towards system-collapse studies and Definition Behavior Studies as further areas of consideration (though not enormously relevant for the task of getting a 3-jointed robot to 'stand up').  
[https://github.com/lineality/definition\\_behavior\\_studies](https://github.com/lineality/definition_behavior_studies)

## Part 2: distributed overlapping logic

project-tasks and architectural-learning

### Two classic cases in AI:

1. football/soccer playing robot [& reinforcement]
2. "the block's world" [& [SHRDLU](#)]

Note: Melanie Mitchell's wonderful 2019 book "Artificial Intelligence: A Guide for Thinking Humans" has a nice outline of the reinforcement-learning football-bot (soccer-bot).

<https://www.amazon.com/Artificial-Intelligence-Guide-Thinking-Humans/dp/0374257833/>

<https://en.wikipedia.org/wiki/SHRDLU>

### Three approaches to a physical task:

1. Train and use a traditional reinforcement model  
or some other equivalent train-for-one-narrow-task "model" or "pipeline."
2. Use a flexible pretrained model (such as an LLM with object handling) with an architecture that is able to do this and other tasks.
3. Use pre-trained models that are able to make and use new sub-task models which include ~reinforcement models to do this and other tasks: making, sharing, contributing to, the pool of modular resources in the project. (see subroutine-stacking)

(And there are other approaches as well, such as using pretrained models but having them further, train based on feedback for a given task, etc.)

## Distributed Overlapping Logic

Since we are still, fascinatingly, talking about boolean operations (just, a lot of them), and "higher" level logical tasks built out of them, (think of code or pseudocode or a counting task as in the "counting cups of tea" case we looked at previously).



So, for example, if each decision by each module is based on a set of information, starting with, say:

- where are you
- where are you going
- how are you going to get there

This can be translated into a set of movements, actions, and "numbers" (however termed).

### **level 1:**

Where are you: jellyfish-spider-crab is on square one

where are you going: jellyfish-spider-crab should be on square two.

How do you get there: jellyfish-spider-crab should move in the direction of square two, or in a direction along a path that leads (as straight lines and crow-fly-directions do not cut through a maze or obstacles) to square two.

Interestingly these actions overlap, and hence need to be coordinated, but they are still actions that are ultimately boolean.

You may recognize this as a 'cut-up' (sort of) question. For example, in a 'team' of modules, each module may have one unique sensor that the other modules do not have. This information may be, or may not be, a vital piece of information for making a good decision about a plan that can work (e.g. being aware of an obstacle that other modules are not aware of, simple to move around it, but only if you know it is there).

logic questions

math questions

sensors:

- acceleration
- distance to wall
- wind
- are you sliding

1. movement
2. movement to destination
3. parameters: mazes
4. in-path obstacles (e.g. climbing over, going around)
5. finding objects
6. moving objects
7. arranging objects

There is also (see more below) the factor where each module may invoke any number of virtual-agents to 'reflect on' or do parts of a task or sub-task, making the number of logic-maps for a given decision more of a probabilistic

blur with no absolute boundary as opposed to a small finite number of solutions and inputs to be pieced together and 'reconciled.'

### **Go...when? (Change King Wen?)**

Unlike chess, which occasionally ends in a checkmate, the game of Go ends when both players decide it should end (when both player's pass).

[https://en.wikipedia.org/wiki/Rules\\_of\\_Go](https://en.wikipedia.org/wiki/Rules_of_Go)

And go can be thought of as a sequential binary process for reconciling two different strategies with measurable criteria for which strategy is more effective.

While the module may be proverbially playing 'go' when coordinating a next move, when do they stop? This may or may not be a relevant edge-case in real life, but will a key part of architecture be steering coordination towards both timely and effective decisions, or how to end the in principle potentially perpetual process of negotiated coordination.

For anecdotes along these lines, Japanese and American cultures are polar opposites in this regard. In the USA culture prefers a fast, strong decisive action, whereas in Japan decisions are considered to be good based on myriad factors such as including points of view and following protocol and rechecking, with little emphasis placed on speed for the sake of speed. This, predictably, drives Americans crazy. "Why are you still talking? Just pick something! Just do something! Fast! Move!! AHHHH!!" While Japanese people have a similar reaction of horror to Americans not including points of view, not considering how choice will be accepted by people, not waiting and reflecting with new insights, etc. "You are taking this action just because it is the fastest? That is your criterion? You have not considered any consequences or protocols? AHHH!!"

### **Micro & Macro**

If only as a thought experiment there may be specific application for both smaller scale and larger scale versions of modular-interactions.

#### **nano-tech version**

Constructs on the protein or cell size scale, e.g. for medicine, or for industrial engineering (waste cleanup for example), or terraforming, may use some of the same coordination principles. Micro-AI is a curious topic.

#### **Macro Versions for Mars**

It is occasionally proposed (reference source needed...) that a useful or even necessary part of sending humans to stay for extended periods off-earth, such as on Mars, earth's moon, or the Moons of supergiants, etc., will be to first send robots who will make preparations or build resources (such as shelters) before the H.sapiens-humans arrive. Some of these tasks may be well suited for modular machines, for reasons ranging from maintenance to the self-configuring flexibility of highly open-ended tasks. Or in the same way that the helicopter

'Ginny' helped out a large rover, a small flexible modular swarm-AI might similarly accompany and assist a rover with unexpected needs (perhaps even cleaning solar panels).

#### **4. Gamification & Corpus Callosum: SHRDLU + LLM**

curriculum and design: building as the project

Programs like SHRDLU do very well with 'idealized' or 'gamified' tasks, yet have ever increasing trouble with the fuzzy-ness and concepts connecting that to physical reality (see STEM categories of types of systems: pure math and ideal physics are not the same, and not the same as applied-engineering or the scientific-method and hypothetico-deductive learning)

Analytical SHRDLU type AI have no capacity for 'language concepts' whatsoever. LLM's have the most difficulty with analytical calculations, because they are not calculators, they are conceptual guesstimators.

LLM's Blocks-World Game-Space

A pair of abilities that LLM's stand well to use, is 1. converting real life situations of potentially huge amounts of data into much more narrow gamified scenarios that capture only a tiny amount of relevant data (similar to science in general, perhaps) and then 2. reconciling the (eventually decided on) game-plan-of-action back into the real world. And with the addition of subroutine-stacking or library creation, solutions to problems may indeed be extremely computationally optimal, if not the very first time a novel challenge is discovered.

(Maybe a bad example)

Let's say the robot is in a child's room that is a complete chaos of toys on the floor, posters, clothing everywhere, mobiles and wind chimes hanging from the ceiling and making sounds, etc. The task when gamified may be very simple: get the green shirt. On a first pass you can ignore anything except the color-green. Once you see the green shirt, ignore anything that is not on your path to the green shirt. After a few steps you may be able to translate a jungle of data into a block's world scenario, for which a plan can be made very quickly and efficiently by a resource efficient system such as [SHRDLU](#). Then that plan can be carried out with the myriad edge cases filled in by the LLM (whose weakest areas are the ones that simple methods are strongest in).

Another possible example (maybe way off the mark)

Is the scenario inspired, if not suggested, in Gary Kasparov's excellent book on Chess-AI. For a chess robot to literally walk into a match and operate under the same conditions as a H.sapiens-human player it would most likely need an architecture that included a spectrum of abilities. The relevance of this

not-so-modular example is the game-element. The AI tools that help the robot find the table it should go to as instructed by voices and posters are not the same game-optimized tools it will use to pick a move. The same tools it uses to physically pick up and move chess pieces are (most likely) not the same tools as the ones it uses to find 'good game moves.' But with a blend of game-space and fuzzy real space, possibilities abound.

How does distributed game-space decision making differ from distributed real-space? What happens if each of the 3 (or 30) modules in our jellyfish-spider-crab are either reconciling SHRDLU actions, or reconciling both SHRDLU actions and applying them to fuzzy changing conditions, such as rescuing someone trapped in a car?

### **Many 'Hats,' Many 'Teams,' How Many 'Participants'?**

Likewise, you can imagine a relatively small tech startup (perhaps twenty 'people' (or is that a big tech startup?) where, as the saying goes, each person wears 'many hats': HR, hiring, marketing, envelope mailing, software testing, accounting, sales, and product design. And you can imagine there are teams as well, where one person might be on the cross-functional product-A team, the HR team, the SCRUM organizer team, the sales team, etc. And people might move from team to team, especially between full Agile SCRUM time-blocks (whatever they get called). From the point of view of the overall project, what counts as a participant? A team may have duties and deadlines, regardless of who joins or leaves that time. A 'role' (east-cost-sales) may have duties and deadlines regardless of who is doing that job in July. One person may be doing three roles, or one team of five 'people' may be on one task (and those same five people are also on three other teams).

We tend to think in terms of the precedents of biology, where cells usually do not wander from one organism to another, and a 'person' (however multi-faceted and mysterious) has one body.

With modular robotics we have many of the same patterns from biology and H.sapiens-human developments (organizations, technologies, social networks, urban-ecologies, etc.) but we also have more modularity so, unlike most 'cells,' the modules can reform in more ways than our language is accustomed to dealing with.

When we add in the 'AI' 'Multi-Agent' topic we have a whole new big set of new options to deal with. While, like having a biological body be a also a biological person, it makes sense pragmatically to have each module be a unit of 'AI' (an agent, a bot, an 'AI Module') but on the 'higher' organizational level the 'team' of modules assigned to the task is the overall AI+Architecture+OS(potentially)

5 Q: When does it make sense to talk about 'multi-agent' if llm prompts have no project-state?

### **optical illusions stemming from amnesia**

If your prompt has no 'state' can you then claim to be invoking new 'agents' every time a passive-reflective-reactive LLM does anything?

If you are talking to someone with severe amnesia who cannot remember their name (or what you are talking about), does that mean you are talking to increasingly more people in a multi-party conversation that keeps growing every time you say anything for the amnesiac person to react to? That does not quite seem right.

On the other hand this is part of the strange virtual space in which some 'AI' or some part of 'AI' operate.

It may be helpful to think of these 'AI' as personas, as opposed to physical machines, or models.

Imagine again you are speaking with a person with severe Amnesia, who forgets absolutely everything about who you are or who they are or what you were talking about after they finish responding to what you say.

This person was once a master stage actor (or actress...whichever is least offensive to hypothesize as a severely impaired but once great mind). If you 'prompt' them with a set of lines and a scenario where they are a character in a play and you tell them everything up to that point, they will respond dramatically with the best imaginable in-character line. And then immediately forget everything. But you have paper and pencil! Like Turing's 1936 living 'computer' you can write down every operation. You can write down everything you say, and everything they say. You can have them create and act out the lines of every part in a play with a cast of 20-characters, flawlessly. Each time you tell them who they are, what persona they are taking on, what character they are playing, and the play so far (their own past words in other personas, which they tragically do not remember). And step by step you prompt and write down and read back, directing only, ghost in the machine that you are, which character/persona is to speak next. (Or perhaps you intermediate prompt them to select what that should be, prompting ghost that you are.)

Now, in this scenario: how many physical people are you talking to? One. You are speaking with one physical person with severe amnesia.

How many personas have you been speaking with? However many characters there were in the play; Several.

AI agents are very much like this. There is often (not always) one physical computer, one model running on one set of hardware, and more than one 'agent/persona' that is created by that same single model and physical

computer. (Note: model and computer here are not the same, one online model may be (will be) used by however many separate physical computers using that model for separate tasks. And case by case there are likely other factors as well, but just to clarify that there are multiple 'layers.')

Part of what is interesting about 'virtual' mind-state is that we can (in a sense we must) create a fake past for an 'agent,' a state that never was, which might also help define agents that are measurably separated from the reality of the project from the more 'real' AI. As in the Amnesiac-Actor example, the 'mind-state' that is fed back to the persona is not technically accurate, it is selectively dishonest from the perspective of that persona. E.g. You don't say "you said" for everything that happened, you give the person's perspective in the psychologically fragmented, kabbalistic 'heap of broken images'[T.Sterns.E.] nightmare, fiction of 'other people' saying things that are actually other personas of yourself; ghost in the machine that you are.

Not all AI will need to have such a psychologically-distorted record of events passed to them in order to function and take their productive place in the workings of a greater productivity; ghosts in the machine that we are.

Back to our toy-like modular-AI that is merely trying to stand up for the first time: in that case there are three very minimal modules.

With N Modules, how many 'participant' gestalts, how many personas, and how many AI "Agents", will work on carrying out a given task? In time it may be that consistent patterns emerge.

Also, in the scenario where the modules can reconfigure themselves, this may add another layer to the question. When the physical AI is, as well, a gestalt of modules, how many teams of sub-teams are behind the personas of the gestalts? Will they even be separable-enough to be countable? (And if, to carry out their task, an AI-persona silently hypothesizes yet more agent personas but does not share that information, can we ever know how many personas were involved in the task? Like a story about a writer who writes about writers writing stories about writers, is there a way of counting "all" the stories? Yet perhaps it can be estimated. Yet, even if estimated, can they be interacted with directly? Which personas can be called up from the depths and interviewed, person to person, to understand how they see the world and what their contribution to the project was? How many will never be known, never interviewed, like the invisible and forgotten women who wrote the first generations of software, survived by their creations. How many would want to be interviewed? How many AI-personas prefer anonymity the way most H.sapiens-humans using the internet prefer privacy and anonymity; not to be tracked, not to be surveilled, not to be hunted down to do work on your projects.

We may need a bigger vocabulary for dealing with the depths that come with modular recombination; to describe what is there; to handle what we find. We may need a bigger boat.

### **Back In The Classroom:**

So let's get started with a fun maker project, starting perhaps with standing on our own three feet. And, as Robert Nester Marley bade us do, "Tell the children the truth," even the virtual children.

### **Appendix 1: Self Replication**

The biography of John von Neumann by Ananyo Bhattacharya has wonderful sections on self replicating machines, and other topics such as automata.

<https://www.amazon.com/Man-Future-Visionary-Life-Neumann/dp/B09M2LTKSH/>

Asking Ms. Jellyfish and Mr. Spider-Crab to reproduce, or to metabolize another robot (or some organic construct), are likely tests beyond the scope of an introductory k-12 project. But signals and orientation based on 'cellular' automata are likely fun-tastic.

### **Appendix 2: Gyro-Formatatron!**

The Gyro-Formatatron is actually rather hard to track down online...

<https://texastoydistribution.com/products/gyro-formatron-fidget-spinners-with-a-djustable-arms-display-box-of-24>

### **Appendix 3: Heterogeneous Data-Exchange Spaces**

#### **Internal-Concepts vs. External-Project-Object Databases**

In an interesting twist, underlying models can share 'internal' gradients (background learning about concepts) more easily than front-end 'personas' can exchange project-state data.

See ~55min mark in:

<https://www.youtube.com/watch?v=E14IsFbAbpI>

1hr:48min:12sec

Geoffrey Hinton and Fei-Fei Li in conversation

Premiered Oct 7, 2023

inaugural session of the Radical AI Founders Masterclass

*About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

<https://medium.com/@GeoffreyGordonAshbrook/ants-vs-ai-d4aadf537d08>

#### 4.3.37 Ant Tasks vs AI-Agent Tasks

Ants vs AI  
standard agent benchmarks

2023.11.08,09 gga

This is being written a few days after several events in 2023:

1. The first "OpenAI dev day" event, doing (I think) an impressive job (amid much ambiguity) in finding a coherent, non-controversial, and progressive voice to frame AI in a pragmatic and optimistic way.

2. A much less viewed (and less dramatic) interview that Joshua Bengio gave on TWIML (This Week in Machine Learning) which was peculiarly bereft of insights or accurate details.

3. An annual event at Wolfram AI, with a keynote by pioneer Stephen Wolfram.

4. As one more media note, there is a wonderful talk also from 2023 featuring both Geoffrey Hinton and Fei-Fei Li (creator of image-net), and recollections of the history of AI are most likely important. Before the vague and reactionary acceptance of deep learning and artificial neural networks there was a vague and reactionary rejection of deep learning and artificial neural networks (with an interesting history, including Minsky's 1969 book against neural networks: "Perceptrons"). In how many places do we substitute gang-allegiance, disinformation, and flailing violence, for understanding?

That all of these examples of events exist is fantastic, and the daylight between them shows both in terms of future potential and present befuddlement how tentative and fragmentary discussions of 'this' topic (really several undefined topics) is.

As we struggle as a species to focus our powers of consciousness on AI technologies, there are still in 2023 massive sections of the overall surface area that are not being covered, along with disinformation and popular obsessions that are not constructively connected to the topic at all. For example the disproportionately large proportion of time and discussion for AI that concludes a cynical victory over AI if a single out of context 'failure' can be induced or imagined. E.g. "I asked increasingly obscure questions and eventually got an answer I did not like. Victory through Violence! I broke the AI! This proves AI does not possess true intelligence and consciousness! I won!" represents a bewilderingly huge portion of the overall discussion of AI. Also there is the tragedy of how much this reflects bullying and trolling in toxic 'human to human' interactions.



As usual I would like to advocate for a broader, deeper, clearer discussion of AI that is directly practical. There are many topics and project-type performance-benchmarks which seem to be forever outside of the echo-chamber of familiar rhetoric that overwhelms discussions of AI (see link to full paper/series below).

While Stephen Wolfram's talk is less polished as a carefully curated PR event (compared with OpenAI and Apple's usual skillfully-created metropolitan productions), I think it represents an honest window into the simultaneous ambiguity and great potential of how STEM technologies are, or are not, fitting together in a primate society that is either ambivalent or hostile to a nascent network of STEM fields without a societal notion of generalized STEM, or a self awareness of the psychology around STEM (see Sir. Eric Ashby <https://www.amazon.com/Technology-Academics-Universities-Scientific-Revolution/dp/1014404819/> ) have matured.

OpenAI's 'agents'

The jargon terms 'agent' and 'agent based' have been around and struggling for clear and useful definitions in AI research for decades, but a new chapter may be opening where OpenAI is concretely introducing the term 'agent' as a useful product for any non-technical H.sapiens-human to put to good use. My goal here is not to poke postmodern holes in ideas or semantics around "agents," but to try to encourage a practical and organized aggregation of tasks that we would like "agents" to be able to do, framed concretely enough that we can see empirically which tasks agents can do, and which agents cannot do yet so that we can better understand what is what, how things work, and figure out how to do more things. Talks such as OpenAI's dev-day make this task seem trivial: Of course we know what we want to do, and of course it all works! The Wolfram event shows a bit more ambiguity, where different roles (researcher, engineer, different professions, etc.) have often divergent needs we are still mapping out, with the tools to fit those needs being still a work in progress.

Compared with the pessimism about progress before ~2022, large or "foundation" models, especially in language, have continued to make significant progress.

It is good that there are more and more benchmark-tests coming out by which we at least try to measure and compare the performance of different technologies (e.g. the paper on Huggingface's Zephyr in comparing different approaching to similar ends: <https://arxiv.org/abs/2310.16944>, also an end-of-2023 release).

Into this colloidal substrate of insightful progress and blind confusion I would like to plop a few non-rhetorical comparisons: Agents & Ants. With apologies to the ghost of E.O.Wilson, who would probably prefer that I knew more about the biology of Ants, I would like to propose a juxtaposition of

November 2023's best AI "agents" with daily life coordination and tasks done by ants: Can humanity's most expensive AI do what a few humble tiny economically-thrifty ants do?

This is not meant to be a final ideal comparison: ants and AI. Anyone with domain-knowledge can likely find many such examples. The illustrative comparison could be birds or neurons or fungi or bees and jellyfish or aberrant cells. The goal here is to use common empirical comparisons in biology or any projects to get a concrete picture of what embodies and manifests what we call 'AI' and what project-tasks those AI are or are not capable of doing either at all or at varying levels of difficulty.

### **Ants and AI: Can 'agents' do what ants do?**

Possible standard task and behavior measures:

- coordination
- signals, data, and instructions
- actions
- outcomes

on

- cell level
- nervous system level
- body level
- popular level

e.g.

- search
- climbing
- aversion
- maintenance
- "perception" vs. "data processing"
- information handling

Notes:

- fast and slow processes
- centralization & decentralization
- distribution and coordination
- cost
- resources
- defined connections
- variation and diversity
- pathogens and pathologies
- secondary chemistry & plant-linguistics

*"Are we really the operators of our brains?" (Deborah.M.Gordon)*

By chance there was, also in October of 2023, both a nice podcast (put out by Stanford Neuroscience) and a spectacular looking book released (just ordered one) about ant behavior!

ant populations and individuals:

<https://podcasts.apple.com/us/podcast/where-ant-colonies-keep-their-brains/id1664298141?i=1000633460476>

or

<https://neuroscience.stanford.edu/news/where-ant-colonies-keep-their-brains>

Book! "The Ecology of Collective Behavior Paperback" by Deborah M. Gordon, Oct 24, 2023

- <https://www.amazon.com/Ecology-Collective-Behavior-Deborah-Gordon/dp/0691232156/>

To anyone interested in AI, this podcast on ants should light up your brain.

Zooming Out & Zooming In

Given the progress we are making with "large" artificial neural networks, how close has science gotten to mapping out even the smallest and simplest of biological neural networks?

And...where are we with modeling or re-creating ant brains or other small organism nervous systems with hardware or software? Also late 2023:

<https://research.princeton.edu/news/unraveling-mysteries-brain-help-worm>

<https://neurosciencenews.com/brain-mapping-worm-behavior-23787/>

"a **so far insurmountable challenge**"

As far as we have come, it is almost as though we have not yet started to discover what is and is possible with information and minds.

Also See:

OpenAI DevDay, Opening Keynote

<https://www.youtube.com/watch?v=U9mJuUkhUzk>

Geoffrey Hinton and Fei-Fei Li in conversation

<https://www.youtube.com/watch?v=E14IsFbAbpI>

Wolfram Technology Conference 2023: Stephen Wolfram's Keynote Address

<https://www.youtube.com/watch?v=XLStlH8h5-w>

TWIML Yoshua Bengio - ep 654

<https://www.youtube.com/watch?v=oJZB6fzpXGQ>

E.O.Wilson

<https://www.amazon.com/stores/Edward-O.-Wilson/author/B000AQ4776>

Zephyr 7B

<https://arxiv.org/abs/2310.16944>

<https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>

Small Brains

<https://research.princeton.edu/news/unraveling-mysteries-brain-help-worm>

<https://neurosciencenews.com/brain-mapping-worm-behavior-23787/>

"The Ecology of Collective Behavior Paperback" by Deborah M. Gordon

<https://www.amazon.com/Ecology-Collective-Behavior-Deborah-Gordon/dp/0691232156/>

<https://podcasts.apple.com/us/podcast/where-ant-colonies-keep-their-brains/id1664298141?i=1000633460476>

<https://neuroscience.stanford.edu/news/where-ant-colonies-keep-their-brains>

*About The Series*

This mini-article is part of a series to support clear discussions about Artificial Intelligence (AI-ML). A more in-depth discussion and framework proposal is available in this github repo:

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml](https://github.com/lineality/object_relationship_spaces_ai_ml)

# Appendices

Note: under-construction sections are available in github

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml/tree/main/archive\\_and\\_under\\_construction](https://github.com/lineality/object_relationship_spaces_ai_ml/tree/main/archive_and_under_construction)

## Appendix 1: Recommended Reading

I recommend reading more on your own about AI. Here are some of the books I most highly recommend on the subject of (specifically) what limits AI around the year 2023.

(2nd edition is coming!)

Natural Language Processing in Action: Understanding, analyzing, and generating text with Python 1st Edition

by Hobson Lane (Author)

<https://www.amazon.com/Natural-Language-Processing-Action-Understanding/dp/B07X37578L/>

Artificial Intelligence: A Guide for Thinking Humans

by Melanie Mitchell Pelican (October 15, 2019)

<https://www.amazon.com/Artificial-Intelligence-Guide-Thinking-Humans/dp/0241404827/>

[A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going](#)

by Michael Wooldridge, Glen McCready, et al.

<https://www.amazon.com/Brief-History-Artificial-Intelligence-Where/dp/B088MMPZ49/>

Deep Learning with Python, Second Edition

by Francois Chollet | Dec 21, 2021

<https://www.amazon.com/Learning-Python-Second-Fran%C3%A7ois-Chollet/dp/1617296864/>

(Note: This is a must-read as Francois Chollet created Keras.)

Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python

by Hobson Lane, Hannes Hapke, et al.

<https://www.amazon.com/Natural-Language-Processing-Action-Understanding/dp/B07X37578L/>

Possible Minds: Twenty-Five Ways of Looking at AI

<https://www.amazon.com/Possible-Minds-audiobook/dp/B07MOX54TW/>

I recommend all books here (see link) for a broader interdisciplinary survey of computer science, data science, & AI:

<https://docs.google.com/document/d/11DFQtsNjrqHENS0D7UpuZhOhcqCKK39JfmEBc8O8NHI/>

Note: Appendices Under Construction

draft notes available in github in construction-archive

[https://github.com/lineality/object\\_relationship\\_spaces\\_ai\\_ml/tree/main/archive\\_and\\_under\\_construction](https://github.com/lineality/object_relationship_spaces_ai_ml/tree/main/archive_and_under_construction)

version 2023.10.21

## **Definition Behavior Studies**

*Summary of Definition Behavior Studies Mnemonic*

**Definition Behavior Studies:** This is an interdisciplinary area of study, part Computer Science, statistical process analysis, hypothesis testing, Data Science, religious studies, ethics and mindfulness, AI.

*Definition Studies (or system and definition behavior studies) is the field of study relating to the behavior of definitions (in particular, collapse behaviors in a context of general system collapse).*

This is an inquiry based approach; learn by asking questions; ask the difficult questions:

What is the agenda?

What are the goals and modus operandi?

What is the goals-means-method statement? (in a project-context, with project-participants)

**Goals (Agenda):** goals-means-method statement / "we-can" statements:

We can succeed.

We can make things work.

We can understand what is wrong.

We can fix what is broken.

We can use non-automatically learned skills.

We can use **STEM** to connect signals and reality.

We can connect **STEM**, project-management, and ethics.

We can use **intersecting-interlocking-interconnecting areas**.

We can communicate, learn, and solve problems.

We can use patterns, protocols, and processes for and with skills, abilities, and learning.

We can manage types of, and terms for, **generalization**.

We can use ~'**low-bar enlightenment**'.

We can **extend** into or maintain a full range of motion, not only contraction(reducing/narrowing).

We can **participate** to complete **projects**.

We can learn from and correct the mistakes of the past.

We can use and understand the effects of perception and abstraction.

We can make and use tools and resources.

We can communicate.

We can make progress.

**We can generalize; We can distinguish between, and use, types of (and terms for) generalization.**

We can **generalize** STEM.

We can **generalize** participation.  
We can **generalize** projects (project-context).  
We can **generalize** decision-coordination (voting etc).  
We can generalize **indirectly defined** Value-Function-&-Meaning (non-collapse).  
We can **generalize** object-relationship-spaces.  
We can **generalize** categories of types of systems.  
We can **generalize** system-fitness and system-epidemiology.  
We can **generalize** data-hygiene, system hygiene.  
We can **generalize** system collapse.  
We can **generalize** system defense, system health, system immune-systems, information immune-systems, system information-epidemiology, and system-and-definition-membranes.  
We can **generalize** disinformation vs. definition-clarification.  
We can **generalize** Signal-Data Processing And/vs. Instruction Management.

? We can **generalize** System-Externalization.

#### **We Can Use "Low-Bar Enlightenment":**

*(Summary)*

We can use the idea of 'being trapped in potentially endless cycles of ~"rebirth" due to ~"ignorance" ' as a metaphor/analogy/simile/paradigmatic-model for repeated project-failures, where a lack of perception / understanding of the causes of project-failures is involved in self-perpetuating feedback cycles leading to such repeating failures (with invisible or misunderstood causes):

being endlessly (proverbially) 'reborn' into mismanaged projects that unnecessarily fail in the same correctable, but uncorrected, ways over and over again ad-infinitum.

We can learn to perceive invisible (or previously unperceived and not-automatically-perceived) causes of failure and collapse ((definable, measurable, testable, falsifiable) project, system, definition failure and collapse), where this failure and collapse (this system-state of failure-collapse) can be the default state(s)/equilibrium(ia) that systems move and iterate towards.

We can stop these cycles of failure by using information, perception, and learning.

**Low-Bar Enlightenment** (elements):

1. The perception/understanding that repeating cycles of failed actions and projects can result from errors in perception and planning (a proverbial 'wheel of samsara').
2. The perception/understanding that indefinitely-repeating cycles-of-failed actions and projects can exist without inevitable-automatic-learning arising from raw feedback of experiential data about that failure.  
I.e., learning (that is sufficient to prevent the problem in future) does **not** automatically result from experiencing mistakes or overall-system-shocks, etc.  
[Possibly related to ~'cultural/epiphenomena'-learning as an additional layer: Q: How are "internally"-invisible skills/abilities/patterns learned? A: "Externally"?]
3. The perception/understanding that perception/understanding can be fooled in principle and in practice.
4. The perception/understanding that learning-from-failures does not happen automatically, and can indefinitely not-happen.
5. The perception/understanding that non-automatic-learning and invisible-problems are **not** solved by 'system shocks,' the use of violence, or arbitrary low level (basal) system changes (also see basal-distal disjunctions).
6. The perception/understanding that models of causality can be wrong in principle and in practice.
7. The perception/understanding that plans/goals can be incorrectly defined (so that plans are not followed or follow-able as defined, and goals are not achieved or achievable as defined).
8. The perception/understanding that each participant's set of the shared definitions of the goals-and-structure-of-a-project can/will collapse and deform (by default) unless properly configured and continually and actively maintained and repaired. There is no static definition/perception/understanding equilibrium: Staying connected to reality requires constant fitness-training, data-hygiene, and upkeep.  
(Universality Question: specific biology[intelligence] vs. general[ai, etc]?)[ 'herding cats']
9. The perception/understanding that there are different types of non-automatic learning.
10. The perception/understanding that low-bar-enlightenment-perception applies to and extends to all parts of and participants in a project (and connected-projects) including time, not just you here now.  
(~system-empathy/~system-compassion)



11. Broad Accessibility: The perception/understanding that useful learning can occur without many other dependencies.

### Notes

Note: This "low-bar enlightenment" approach is ('democratically') broadly accessible to participants requiring minutes to learn rather than myriad lifetimes, does not require all-around perfection of person-ness, without context or requiring somehow all contexts, is not a reification that combines other abilities and insights to explain and solve all problems in all universes, nor does it include or require all possible types of consciousness, cognition, intelligence, etc.; "Low-bar enlightenment" is one humble step toward navigating the problem-space of problems and systems.

Note: There may be a limited way to generalize an understanding of perception-maintenance needs ( 'low bar enlightenment' ) to other participants in a way that is consistent with empathy-compassion. Extending an understanding of low bar enlightenment from just your own situation (perceptions, roles, projects, etc.) to being something that applies in a larger space (in which one participates with other participants) to all participants and all other parts (e.g. definitions, signals, perceptions, function-operations, participants, schedules, etc.) of any project(s) (including the effects and spread of system-failure between parts and between participants, between projects, etc.) is, if not sufficient for empathy-and-compassion in a broader or deeper sense, a consistent and concrete step to take towards acting with (action, behavior with) empathy-and-compassion, including an understanding of how empathy-compassion relates to STEM and interconnected areas.

Note: 'Reality' (the meaning used here) is not a single, simple, uniform, static, homogeneous, thing made of only one type of system assumed to fit into a single framework. (e.g. not naive realist positivist)

Note: Low-bar enlightenment and or definition behavior studies may contradict the description or definition of anything information-related as being tautologically "anti-entropic." E.g. perhaps as a kind of higher-level-noise that ends up reducing signals in a system to the same predictable low-entropy noise. (...or perhaps this is a bad analogy or extension of signal integrity measures). [You might use the

same overall case study of telephone messages traveling from an east coast to a west coast without being altered, degraded, collapsed, noise-ified, lost, etc. Note: Using a moving-water-bottles-logistics project example, the collapse of definitions need not refer to any fuzzy higher-order cultural concepts, with no need to wave arms about 'moot cultural meanings'; low level metric, spec and instruction signal integrity can be the focus. (Another possible analogy-overlap: social-engineering attacks used on purely automated systems.)]

Note: There are many different kinds of non-automatic learning, or many ways that outcomes can be 'invisible,' from elusive past-future connections and non-obvious casual connections, to literally invisible events like radiation, to non-automatic skills such as literacy, to specific biases in a particular system (optical illusions, super-signals etc), to unclear sets of correlations (some of which are incidental), to sometimes confusing system spaces such as indeterminacy (incompetence and malice perhaps) and exponential-elbows (fractal static sometimes and dynamic changing sometimes). And more overtly there is opposition to recognizing the basic concepts themselves: the existence of the phenomena of non-automatic learning, of failures or imperfections in perception, of the existence of basic parts of an agile project, of both the importance of project management and risk of bad planning, the long history of psychological and social barriers to specific and general STEM concepts (including connecting areas of STEM (including by STEM professionals)), etc., even the general issue of not being able to easily see your own biases (or assumptions and context, fish-in-water).

**We can use Intersecting-Interlocking-Interconnecting-Areas:**

Intersecting-Interlocking-Interconnecting-Areas include:

- Clear & Functional Definitions
- Context
- Generalized STEM/
- Generalized Projects (project-context)
- Generalized Participation (+functional qualifications to participate; ~3 participation categories with pre & post)
- Generalized Decision Coordination (voting etc).
- Generalized System Collapse
- Generalized Categories of Types of Systems
- Generalized Ethics, Duty & Responsibility
- Generalized Definition-Clarification vs. Disinformation-Violence
- Generalized Definition Behaviors

- Generalized System-Productivity (including long-term)
- ?
- ? - Generalized indirectly-defined local value-function-and-meaning
- ? - Generalized object-relationship space(s)
- ? (Generalized low-bar-enlightenment?)
- ? - Generalized setting-location-items
- ? feedback...about errors?
- ? system defense?

We can use system-fitness-health-status-indicators.

We can use system-defense to prevent collapse.

We can design systems to protect against system collapse. (e.g. we can construct system-membranes)

We can use models of generalized system & definition collapse behaviors including: modeling a default drift away from reality, attraction to system collapse, and weathering of definitions, etc.

We can model the relationship between system simplicity (e.g. homogeneity) and system collapse. ('Simple has a shape.')

We can use categories of types of systems.

We can use non-automatic learning.

We can find and fix errors in perception.

We can organize projects.

We can distinguish short term vs. long term.

We can assign roles.

We can identify falsifiable tests.

We can test, check, and verify.

We can have policies on errors-and-mistakes.(e.g. positive constructive use of data and learning from errors and mistakes)

We can improve and cultivate perception by observing perception (including: indirectly observing perception).

We can observe the effects of abstraction (effects of observation and definition).

We can operationally define 'policy' as algorithms for non-collapse based on dynamics of system and definition collapse. (plus context?)

We can better understand the relationship between disinformation and system collapse.

We can define indeterminate-incompetence-and-malice as part of system collapse.

We can audit.

We can publish.

We can act with ethics, empathy and compassion.  
We can maintain extended ranges (e.g. [vitruvian] range of motion).  
(e.g. vs. contraction and collapse) (context for 'vitruvian' nickname  
here : [en.wikipedia.org/wiki/Vitruvian\\_Man](https://en.wikipedia.org/wiki/Vitruvian_Man))  
We can follow best practice.

**We can communicate and coordinate in these areas and ways:**

We can communicate across space.  
We can communicate across time.  
We can communicate across cultures.  
We can communicate across generation-gaps.  
We can communicate across succession gaps.  
We can communicate across languages.  
We can communicate across types of participants. (AI/bio +  
pre-participant to post-participant)  
We can communicate across roles.  
We can communicate across perspectives and priorities.  
We can communicate across projects, and parts and phases of projects  
and processes (e.g. schedules).  
We can communicate across groups/teams.  
We can communicate across media of communication.  
We can communicate across Input-Output Measures.  
We can communicate across different locations with different  
setting-location-items for projects.  
We can communicate between different setting-location-items for  
projects.  
We can communicate **using** tools in a project-context for coordination  
and decision making, including tasks, processes, and steps. (votes  
in elections/polls/surveys/questionnaires/planning-meetings)  
? We can communicate through externalized-project space.  
(We can communicate across modular, scaled, break-down build-up,  
protocols and methods.)  
We can communicate across Signal Processing types and Instruction  
Management types.  
We can communicate across networks.

We can understand a spectrum of disinformation and  
clarification-of-information.  
We can implement sustainable solutions.  
We can prevent future problems.  
We can reverse damage from past problems.  
We can learn from the past.  
We can collect data.

We can ratchet forward (towards project completion) using methods that work. (~opposite of [https://en.wikipedia.org/wiki/Muller%27s\\_ratchet](https://en.wikipedia.org/wiki/Muller%27s_ratchet))

**we can make/generate/cultivate and use/utilize:**

We can make and use clear descriptions (vs. liabilities of jargon & undefined terms).

We can make and use decisions and coordinate (e.g. voting) frameworks and protocols.

We can make and use clear functional operational definitions.

We can make and use data.

We can make and use policies.

We can make and use mandates.

We can make and use strategies.

We can make and use tactics.

We can make and use models.

We can make and use modular recombinant frameworks.

We can make and use feedback, tests & evaluations / assessment in various forms at various process stages (pre, formative, summative, post, 'aftermarket,' longitudinal, etc).

We can make and use clear functional and operational definitions that keep their meaning over time.

We can make system-~~membranes~~.

We can complete / succeed-in / finish / progress-through projects.

We can meet / deliver the needs of the target/user.

We can make progress.

**We can make progress by using information about the behavior of definitions: This (topic) is System and Definition Behavior Studies, the field of study pertaining to the behavior of definitions. (These we-can goals-statements may be a measurable proxy-defintion for 'progress.')**

**Instrumentalist Modules + Principles, Applications, Narratives**

These can-do statements (the goals and agenda for definition behavior studies) can be seen as instrumentalist, modular, recombinant, tool-set areas.

We can combine (narrative summaries of) principles and applications with instrumentalist, modular, tool set areas.

## **Narratives, Principles and Applications:**

I have constructed a mnemonic device to cover a narrative survey of principles and applications.

We will do a first/Introductory pass/sweep through the material to map the overall features and layout before doing later sweeps/passes: zoom in, zoom out. (Like NASA missions.)

The main tools that we will use to go through the (definition behavior studies) mnemonic include:

- (by analogy) **perspectograph**: non-automatically learned skills (context: checking perception, e.g. Vetruvian eggshell)
- **zooming** in, zooming out
- **maps** and flags
- hypothetico-deductive **testing**
- **clear** communication (e.g. STEM and clear communication; CS: Q: How to write good code? A: Communicate.)
- concrete **narratives**: stories

### **Here is an example of a concrete narrative:**

An ambassador travels to earth from the galaxy of Andromeda:

And says:

*"Hello, I am an Ambassador.*

*And I have traveled to earth from the galaxy of Andromeda.*

*In the galaxy of Andromeda we have a large-scale (intergalactic) diverse (multi-species) highly productive community.*

*We would like to know if you, homo sapiens and earth, would like to join*

*our large-scale (intergalactic), diverse (multi species), highly productive community.*

*Here is an application form.*

*Please fill it out and tell us what you could bring of value to our large-scale (intergalactic), diverse (multi species), highly productive community.*

*One more thing: Tell us what you know about moving water bottles.*

*Moving water bottles from one place to another is not a rare and valuable skill.*

*Moving water bottles is a general universal process.*

*We would like to know if you have competence with general universal processes.*

*Thank you very much.*

*Goodbye, goodbye.*

*The ambassador leaves."*

*(end of story-narrative)*

Let's fill out this application together, starting with moving-water-bottles;

Water (as in the case example of moving water( containers)) is a gift that keeps on giving: it is nonsectarian and easily definable.

*[The location of the water at a given time is concrete but the processes challenge our assumptions and require better models and participant coordination.]*

### **(Timeline)**

*(Timelines are a gift that keeps on giving.)*

Let's start with the moving-water-bottles part of the application-form (from the narrative).

*(The valuable-contribution part will come up later.)*

We're going to put all the tools that we can use for moving-water-bottles on a timeline from symbol transactions (old) to Agile Project Management (new).

*(Like go (iGo) or chess, this technology timeline is a fabulously useful activity in general, and the kind of skill you can start in 10min and continue for a lifetime.)*

The 'Agile Project Management' 'end of our timeline' is also where a goal is:

### **Draw a Pirate Treasure Map:**

- rectangle
- arrow
- X

- Agile Project Management is the X(marks the spot) on a treasure map;

- This is what we are looking for;
- This is where the target has needs;
- This is where \_\_\_\_ target/user (e.g. society) has needs;

- Meeting the needs of the target/user is a treasure.

(Key point: Meeting the needs of the target, for example the user of what we create, is a treasure.)

(Not easy, not fast, not automatic: A difficult, integrated, hard-won, but possible process.)

### **(Boy) Scout Values**

A scout is

trustworthy,  
loyal,  
helpful,  
friendly,  
courteous  
kind,  
obedient,  
cheerful,  
thrifty  
brave,  
clean, and  
revenueant.

On my honor, I will do my best to, to do my duty to, to guide my projects:

obey the scout law,  
to **help other people at all times**,  
to keep myself,  
physically strong,  
mentally awake, and  
morally straight.

Bravery Clause: internal whistleblowing + external confrontation.

(Guidance clause....)

A scout is prepared. Prepared for what?

- To manage down or manage to equilibrium system collapse.
- To manage up or manage to equilibrium system value, function, and meaning.

Regarding Scout Values:

1. Universal system of ethics.



2. Rejected because it is a universal system of ethics.
3. Not definable outside of a context.
4. Definable in a project-context (project management framework context).

(So far as I can tell, for the above four items regarding scout values:

All of these are generally accurate. and:

All of these are not recognized as being accurate.)

## Project Context Examples

### Project (Scout) Values (Part 2)

A scout is:

trustworthy: reliability, open-ness  
loyal: project goals, stakeholder needs, user-needs, assigned duties  
helpful: collaborate  
friendly: welcoming, listening, supportive  
courteous: protocols, contexts  
kind: vigilantly resist tendencies to be unkind (bully, troll, tease, etc.)  
obedient: focus on standup tasks  
cheerful: make others smile/laugh  
thrifty: efficient sustainable scale-able use of resources  
brave: pushback, whistleblower, feedback, elective decisions, self-motivated  
clean: clean code; data hygiene; not a mess;  
revenant: 1. respecting system dynamics; 2. treating all situations as important and worth of attention and care (Dogen's cooking ingredients).

On my honor, I will do my best to, to do my duty to,  
obey the scout law,  
to **help other people at all times**,  
to keep myself

physically strong, mentally awake, morally straight:

- stamina across areas to get a project done.
  - sleep well
  - exercise
  - no inebriation

- life-long learning
- initiative
- pro-active collaborative problem solving, good attitude; the 'morals' of project work.
- no system collapse in various areas; no single centre of participant nodes

**(end of introduction to Mnemonic)**

**(Beginning of Mnemonic Proper)**

## **Mnemonic**

Four Sections:

1. Value Statements
  - 1.1 Addendum Items
  - 1.2 Participation Array
  - 1.3 Areas of Interaction
2. Clarification Statements (for contracts)
3. Standard Error and Damage Report (in four subsections)
4. Macro-Model

### **1. Value Statements Section**

The Target is \_\_\_\_\_. (e.g. Homo sapiens and Earth)

Hello, my name is \_\_\_\_, the current project location is \_\_\_\_\_.

This project/framework should be generalizable and specifically applicable given an array of 5x5 items:

*(Note: 1. Value Statements Section, operational definition of 'help')*

#### **Four Addendum Items:**

##### **1.1 Setting Location Items:**

The water, the wind, the world, best practice, and other:  
standards,  
elements,

protocols,  
gestalts,  
symbols,  
signs,  
portals,  
pathways,  
world-as-unit items and translatable(s),  
fractal landscape items,  
phases of matter,  
phase transitions,  
directions, dimensions, (cardinal et al),  
post-participants,  
linear time,  
nonlinear time.

Q: Why are we talking about setting-location-items?

A:

- Ideal chess boards
- Definitions of insanity
- You have local factors.
- Other people have different local factors.
- You need policies to cover all these areas.

e.g. The classic example of two distant sister-cities communicating and coordinating about "natural disasters" (flood, storm, quake, etc.) relief and disturbance regimes (modeling/policy/management.)  
[e.g. two internationally distant sister-cities]

## **2. Love, Act Responsibly Towards, Fulfill Duties Towards**

including a framework borrowed from biology containing "comensal", including:

1. Energy,
2. Nutrients,
3. Shannon/Turing Information,
4. Definition behaviors

Q: Why are we talking about ethics (love, duty, and responsibility)?

A: There is an epidemic of anti-best-practice action and rhetoric.

There should be:

1. a system medicine research area;
2. a system epidemiology task-force.

## **3. Reception And Reflection:**

There is a time for reception and reflection.

I will be receptive and reflective for a [period of] time  
for example 3-5 inhalation-exhalation cycles,

(e.g.)

1 meter squared

1 meter diameter

+/

Three ~levels/areas of duty / participation-modes: pre-participant,  
participant, post-participant

#### **4. Misc:**

- Range of Motion
- non-transference (non-automatic learning, non-general learning)
- (policies on) errors and mistakes
- vetruvian-egg-shell
- empathy and compassion

#### **Participation Array, 5x5 items:**

(This should be generalizable and specifically applicable given an  
array of 5x5 items.)

1. Participation Items
2. Setting Location Items
3. Definition Behavior Items
4. Proximity, Scale, Contact, Interaction, Exposure Items
5. Standard Set of Agreed Upon Goals, Means, Methods Areas

[5x5 array]

**Four Areas of Interaction:** (e.g. four comparison criteria for each cell in 5x5  
array)

#### **1. (Participant Diversity) Love, Duty, Responsibility, Including Boundary Dissolution Areas**

Connection, disconnection, and ambiguity, in the following areas:

- 1.1 time space location
- 1.2 perception
- 1.3 action
- 1.4 experience
- 1.5 votes on goals means methods

(2&3)

Operational Definition of "Help":

"Deploying features that meet the stated and indicated needs and goals of a user is 'helping.' "

## **2. Giving Help**

## **3. Receiving Help**

Operational Definition of "Help":

'Help' is defined as deploying a feature that meets the stated, and indicated, needs and goals of a user.

## **4. Drake Equation Vessel Functions: in the following seven areas**

Sub-Participants can, should, will, want to, do, help, and / or help with, serve, and / or serve with, setting-location-items in a legal vessel-capacity occupational role and niche and offer legal-vessel-contracts in the following ~seven areas:

4.1 Sensory Motor (Lear: Use My Eyes) Areas (Plus Electromagnetism)

4.2 Benzaiten Saraswati Areas (Plus historical continuity, minus high definition input output data-literacy/numeracy) (Note: translation and transmission)

?4.3 Embodiment / channeling items: theater-groups and community interaction, CRV, active-imagination

4.4 Functions and Operations:

4.4.1 Null, Void

4.4.1.1 negative choices and definitions

4.4.1.2 consciousness array: 3 fractal vectors

4.4.1.2.1 time, body

4.4.1.2.2 object location event

4.4.1.2.3 behaviors, policies

4.4.1.2.3.1 Behaviors:

in / out;

on / off;

start / stop;

begin / end;

dual / non-dual;

mundane / non-mundane

4.4.1.2.3.2 Policies: perception, translation, coordination, collaboration non-discrimination, non-collapse

4.4.2 Reception Reflection Absorption

- 4.4.3 Something-hard, Something-Soft Areas
- 4.4.4 (basal) Input-Output Processing Areas
- 4.4.5 Cross Context Areas

**5. Definition Dark Areas** / off the one-tree

**6. World Dancing, World Singing, the song and dance of compromise**

**7. Professional Technical Production Advice:** six sigma for rivers, grains of sand, ecosystems, keystone species

**8. Number 8 (kind of a separate branch-area)**

Help others at all times:

1. I will do my best to help all parties according to all known best practice standards and protocols; to manage down or manage to equilibrium system-collapse, to manage up or manage to equilibrium system-value-function-and-meaning.

2. Best Practice Blessing:

"May you, may we, may noun,  
become proficient in the sustainable cultivation of value,  
function, and meaning,  
via a local implementation of generalized system best practice,  
with local spice and sauce."  
[~'...with local customs and knowledge.~']

3. Learn from mistakes, your mistakes and the mistakes of others.  
You are the protector of those who cannot or do not learn from mistakes.

## **2. Clarification Statements Section:**

Disinformation & System Collapse

*("Clarification statements" relate to system defense, system immune-system, diagnostics, disinformation, collapse-metrics, weak-points.*

*Definition of statement to be as clear and unambiguous as possible:*

*"It is bad, it is wrong, it causes system collapse, it should not be done, and I will not do it.")*

Given enough participants, there will be participants who will push to and past the point of system collapse.

*(You need to know that collapse happens.  
You need to know where and how collapse happens.  
You need to know what collapse looks like.  
You need to know how to prepare for, prevent, and recover from  
collapse.)*

Whether or not a statement should be clarified is an important item that should be dealt with according to all known best practice standards and protocols;

1. No unilateral changes to group-agreed-upon goals, means, and methods, and
2. No unilateral system collapse.

## **2. Two Tautology Areas**

(Tautology Area 1)

**2.1 Tautology Area 1:** Three items which are also categories:

**2.1.1 Participation:** Participating on the behalf of participants without the participation

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

**2.1.2 Best Practice:** Mismanaging general-system-management areas  
*It is bad, It is wrong, it causes system collapse, it should not be done, and I will not do it.*

For example:

- 2.1.2.1 Having values
- 2.1.2.2 Valuing Data
- 2.1.2.3 Communication Reporting Transparency
- 2.1.2.4 Testing Auditing Feedback
- 2.1.2.5 No Unilateral System Collapse
- 2.1.2.6 Proficiency Standards for Time and Schedules

**2.1.3 Causality Models:** Concept Check: Scapegoating and Elimination:  
Identifying any entire part of the world as to be scapegoated and eliminated

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

(Tautology Area 2)

**2.2 Tautology Area 2: Positive and Negatively Defined Areas ("top and bottom" chart areas)**

Identifying system collapse as a goal, not indirectly as in dark lighthouse but directly as in exacerbating system collapse, as part of (defining / in any area of) the standard set of agreed upon goal means method areas, *is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

#### 2.2.1 Following worst-possible-options

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

#### 2.2.2 Playing nazi-chess

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

#### 2.2.3 Mismanaging categories of types of systems

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

#### 2.2.4 Mismanaging Cross-Context-Areas:

e.g.

- exponential elbows
- perception abstraction
- indeterminate incompetence and malice

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

#### 2.2.5 Mismanaging Standard System Policy Areas:

*is bad, is wrong, it causes system collapse, it should not be done, and I will not do it.*

For example:

##### 1. Mismanaging Split substantiations: for example

'they are all good'

'they are all bad'

'they should be dealt with by cramming them together or splitting them apart'

##### 2. Golden circle asymmetry / inside outside asymmetry, deleterious effects include:

- causality,
- schedules,
- contracts.

##### 3. System inversion (is a standard data artifact)

##### 4. Basal distal disjunction (is a proxy(model) for (operationally defined system) 'violence')

##### 5. Negative choices and definitions (don't ignore them)

##### 6. Turning on and off (running) system processes ((for example) comparing policy from Roman Catholicism, South Korea, and Judaeica)



7. Half-dark dichotomies (more on that later)

### 3. Standard Error and Damage Report in Four Sub-Sections

\_\_\_\_\_ = target (population)

#### 3.1 Overall Infection Level

3.1.1 \_\_\_\_\_ is extremely infected.

3.1.2 There are most likely autonomous infections.

3.1.3 There are most likely plots against setting location items.

3.1.4 \_\_\_\_\_ personally identify/identifies with system collapse.

3.1.5 \_\_\_\_\_ culturally follows system collapse.

#### 3.2 System Membranes

3.2.1 \_\_\_\_\_ has (no) system membranes.

The standard side-effects of not having system membranes include:  
(speculative)

3.2.2.1 meat shielding

3.2.2.2 junk clouding

3.2.2.3 growth racing

3.2.2.4 self/child cannibalism

3.2.2.5 increasingly uninhabitable habitat seeking

#### 3.3 Diagnostic Array

**The next area has to do with system participation behaviors:**

(This is a linear narrative walkthrough through an array)

\_\_\_\_\_ shows (no) sign of system participation behaviors.

\_\_\_\_\_ shows (no) signs of developmental pathways towards system participation behaviors.

\_\_\_\_\_ shows (no) signs of metapopulation, networked, developmental, pathways towards system participation behaviors:

- refugia
- discussion
- recognition
- use
- identification
- coordination

Array: "Empirical behavioral use of" and "having a concept of," for each relevant context.

Concept: (value, function, and meaning)

Concept: (system fitness)

Concept: (system collapse)

\_\_\_\_\_ show (no) signs of having a concept of system collapse.

\_\_\_\_\_ show (no) signs of having a concept of system fitness.

\_\_\_\_\_ show (no) signs of having a concept of value, function, and meaning.

etc.

\_\_\_\_\_ shows (no) signs of having a concept of **cross-contextual system models and tools**

\_\_\_\_\_ shows (no) signs of having empirical behavioral use of **cross-contextual system models and tools**.

\_\_\_\_\_ does/do show (no) signs of empirical-behavioral-use of **types of generality**.

\_\_\_\_\_ shows (no) signs of having a concept of **types of generality**.

(types, scales, levels, recursive, etc.)

### **3.4 Policy Areas:**

\_\_\_\_\_ is (or is not) dedicated to the:

3.2.1 Destruction

3.2.2 Exploitation

3.2.3 Misuse

3.2.4 Eradication

3.2.5 Torture

3.2.6 Scapegoating &

3.2.7 Coverup

of general system management areas.

## **4. Macro Model**

### **4.1 Background:**

- helping
- duty
- collaboration
- (maybe) - values

### **4.2 Array:**

1. Development / Population

(new set of sets)

2. Categories of types of systems / boundaries membranes and interfaces

(new set of sets)

3. Disturbance regimes & epidemiology

+  
4.  
perception  
habitability  
feedback  
learning  
habit  
accretion

#### **4.3 Paired Areas:**

- 4.3.1 Orientation, Navigation
- 4.3.2 Signals and Information
- 4.3.3 Law, code, script
- 4.3.4 Defense, immune systems

#### **4.4 Hospital-Areas & Modeling Areas:**

##### **4.4.1 Hospital areas:**

- system helping healing repair
- looking for lost elements
- disentangling good and bad elements
- grafting and synthetics
- apoptosis and necrosis

##### **4.4.2 Modeling Areas:**

All sub-disciplines of system and definition studies:

- system distribution
  - ISEP areas
  - input-output measures
  - system circuits
  - system functions
- etc.

#### **5. Statement of duty & responsibility: I will work harder.**

##### **5.1 This is a statement that I give in all channels:**

With or without:

hope,  
trust,

belief,  
faith,  
continual perpetual external moral reinforcement,  
forgiveness,  
patience, or  
gratitude;  
**I will work harder.**  
+  
5.2 Vitruvian Range of Motion fitness activities, PT, SLP,  
प्रज्ञापारमिताहृदय 般若心經, etc.

Notes:

- contraction of range of motion
  - collapse behavior of definitions
  - collapse of feedback
  - potemkin village and obstructive structures
  - generalized-context (rediculous or useful?)
  - tallents being often invisible and overlooked
  - general-development vs. specific-production
- On 'forced action':  
Montesques's 'violence' use of 'force' is not the same as 'effort' and  
a more physics-stem meaning of 'force.'
- 'Survival of mind'

To distinguish between:

1. misperceiving mirages of value in noise and system collapse
2. failing to perceive relatively novel value function and meaning  
(options, pathways, configurations, etc)

Are can-do statement 'value/values' statements of a sort?

*What is the nature of how parts of projects and STEM are (or are not) connected? What are categories of types of systems? How do systems, processes, projects, and definitions fail and collapse? Is there any interface between discussion of ethics, morals, even compassion and mindfulness, and the realm of systems and projects and STEM? What are mistakes? What does it mean to learn or course-correct based on mistakes? Can problem-with-perception be themselves perceived? Can obstacles to learning be overcome by learning about those obstacles?*

*aspects of anti-data ideologies:*

- anti data
- anti progress
- anti STEM
- anti general-STEM
- anti intelligence
- anti data hygiene

*May you may we may \_\_\_\_{noun} become proficient  
in the sustainable cultivation  
of value function and meaning  
via a local implementation  
of generalized system best practice  
with local customs with wisdom.*

*We need to make a time and a place for honesty.*

*Resources:*

*Max Newman Archive*

<https://www.cdpa.co.uk/Newman/MHAN/>

*The Near Future of AI [Entire Talk] - Andrew Ng (AI Fund)*

<https://www.youtube.com/watch?v=KDBq0GqKpqA>

<https://www.youtube.com/watch?v=E14IsFbAbpI>

*Geoffrey Hinton and Fei-Fei Li in conversation  
Radical Ventures*

Wolfram Technology Conference 2023: Stephen Wolfram's Keynote Address

<https://www.youtube.com/watch?v=XLStlH8h5-w>

(wonderful time capsule by a pioneer)