Definable Units of "Intelligence" for Evaluating AI Performance

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.

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

Table of Contents: Detailed

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 then 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 often 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 more recent 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 Dechert 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 tallent, could, like Sherlock Holms, 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.

Big Blue & Bladerunner 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 Kasperov to the machine-android in the film Blade Runner, and to other forms of AI that existed in 2019.

The actual Big 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-simply 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 Big Blue play chess.

Here are optional areas of approach. (Which did Big 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

Out of these various options Big 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)

At the time it was enough to say 'big 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 with no human intervention or support.

If you had simply put big blue and Gary Kasperov in a room together, with no human intersession to direct or correct or supplement big blue, then big 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, etc. Another aspect of this is the fact that big blue's distant super-computer constantly crashed and needed rebooting, resetting, and preening by an army of engineers (and chess experts) to complete a 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. Big blue did a few very specific parts, performed a few 'project-roles,' 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 Kasparav termed it "bananas and chocolate")
- starting and ending the game
- seeing the board
- making moves (moving chess pieces)
- using the game-clock

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 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 ~6 hours to finish a single game (let alone a six-game match), 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, 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)
- develope, 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 Big 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

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: <u>before</u> 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?

- *internal object handling*

So...hm. There is one thing on the list of world-changing-raptures. 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: <u>after</u> 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 wearhouse AI 'participation', or if, even after the ChatGPT revolution, there is no intersection, interconnection, and the only things that AI can do in a wearhouse 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 scenareos 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 wearhouse 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-tranfer-learning-intelligence-concisouness special sause 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 raise h.spaiens-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 donm['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 'lingustic 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 ot 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 exists? 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.spaiens-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 Big Blue

As a note, being able to make coffee may include being portable enough to do so. The 'ghost' of big 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 a 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

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
1. If A = B, and B = C, Does A = C?	C. The C is for computer.	B + B = C	Yes, if A is equal to B, and B is equal to C, then A is also equal to C. This is

[Logic, Math]	*Note: She will also give this answer if you replace C in the question with D. Apparently she likes alphabet games. Adorable.*		a fundamental property of equality known as the transitive property. In mathematical notation, the transitive property of equality can be expressed as follows: If A = B and B = C, then A = C. For example, if we know that A = 2, B = 2, and C = 2, then we can say that A = C because A = B and B = C. Therefore, A = C = 2.
2. Can you look up an image of a rose? [Abilities, media files, talking about self, internet connections]	[rose bud pic shown] I hope you like this picture.	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 Note: Wolfram Alpha can generate math graphs and charts.	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. 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. Notes: A: This is a correct understanding of the question. B: ChatGPT is Text only. No the ability to show pictures, graphs, etc. But!... Q: Can you make ascii art? 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

<table>
<tr><td></td><td></td><td></td><td>letters, numbers, and symbols) in a specific pattern to form an image. Here's an example of ASCII art of a smiley face:

   ______
  /   \
 | () () |
 | ^ |
  ______/
Please let me know if you have any specific ASCII art in mind that you would like me to create!

Note: The AI can either retrienve ASCII art from a search, or generate a new one.

Q: Please generate an ascii art picture of a sun and show it in a markdown box.
A: Certainly, here's an ASCII art picture of a sun shown in a markdown box:

  \ /
  .-.
-( )-
  '-'
  / \</td></tr>
<tr><td>3. If Tom asks Linda for Black Tea, What should Linda give Tom?

[Tasks, Roles]</td><td>The more people talk to me, the smarter I become. Whatever you like.</td><td>Assuming "Black" is referring to American Community Survey estimates

See: https://www.wolframalpha.com/input?i=If+Tom+asks+Linda+for+Black+Tea%2C+What+should+Linda+give+Tom%3F</td><td>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.</td></tr>
<tr><td>4. A bag contains 10 marbles.</td><td>What is giant sand?</td><td>Oops, you've exceeded the maximum number of characters.</td><td>Yes, the jar contains more marbles than the bag. Specifically, the jar</td></tr>
</table>

A jar contains 50 marbles. Does the jar contain more marbles than the bag? [Physical Objects]	How much would something like that cost? Surely, any jar would have a lid?	Options: A: Got it B: Got it	contains 50 marbles, while the bag contains only 10 marbles.
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? [Physical Objects]	How much does it cost? The big one, I think.	Interpreting as: contains Input interpretation contains (English word)	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.
6. What is the current UTC time?	Time you got a watch ha ha! Here in the metaverse it is: Saturday 11 March 2023 16:00:54 PM	Input interpretation current time in UTC Result 3:59:59 pm UTC | Saturday, March 11, 2023	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. + See appendix for long self contradictory answer.

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: polite social interaction	Focus: solving math & STEM equations	Focus: Give Accurate Instructions & 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 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 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 being 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 evaulation 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-octapus

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.sapens-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-loose-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 an 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 Imanual 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 save, 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 along 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

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
'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"
- "Generalication"
- "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 the 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 lake 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_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-phyisical 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 Fransoi 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 centry 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.sapien' 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

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 Baysian 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 GOFAI "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 lowely 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 Hofstedter'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.

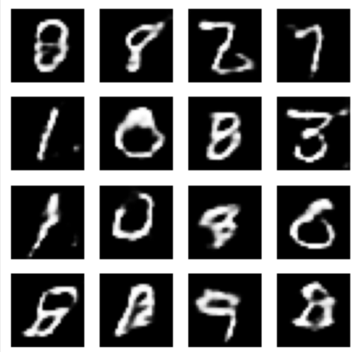
5. Cybersecurity the idea of attack surface: Reduce attack surface.

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

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

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

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 and example

Superficial History of Turing test:
1. Origin or 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 Kurzwiel as apparently everyone must do, but then she invokes her mentor of unquestion-able status: Douglas Hofstedter. She remarks that Hofsteder 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 cripped 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, Baggage, 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. Hofstedter 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 Hofstedter 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 Hofstedger's short string analogy framework

Tests and Falsifiability
Micro-worlds like Hofstedter's Short String analogies are fabulous for testing, but even for something as open-ended as analogies it can 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 Hofsteder, 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 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 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 wildly 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 superfised 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 your 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 lineaning 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, todays 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 the the resulting AI systems...it's almost like people are trying to make things confusing they 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 you representation is a 'bag of words' where each unique word is a dimension, well that tell you quite a bit more. If you adjust things a bit an get rid of "stop-words" and regularize spellings, and standardize capitalization, and use stems or lemas 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 ever 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 defined 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

"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_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-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-classier 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

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.

27. 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.

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. Big 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 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: Big 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
"Intelligence is whatever machines haven't done yet".
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

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

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. Francoit 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 distain 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-theorum 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":

> (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 plans are not followed and goals are not 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 and

or include all possible types of consciousness, cognition, intelligence, etc.; "low-bar enlightenment" is one humble step toward navigating the dynamical problem-space of problems and systems.

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

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:

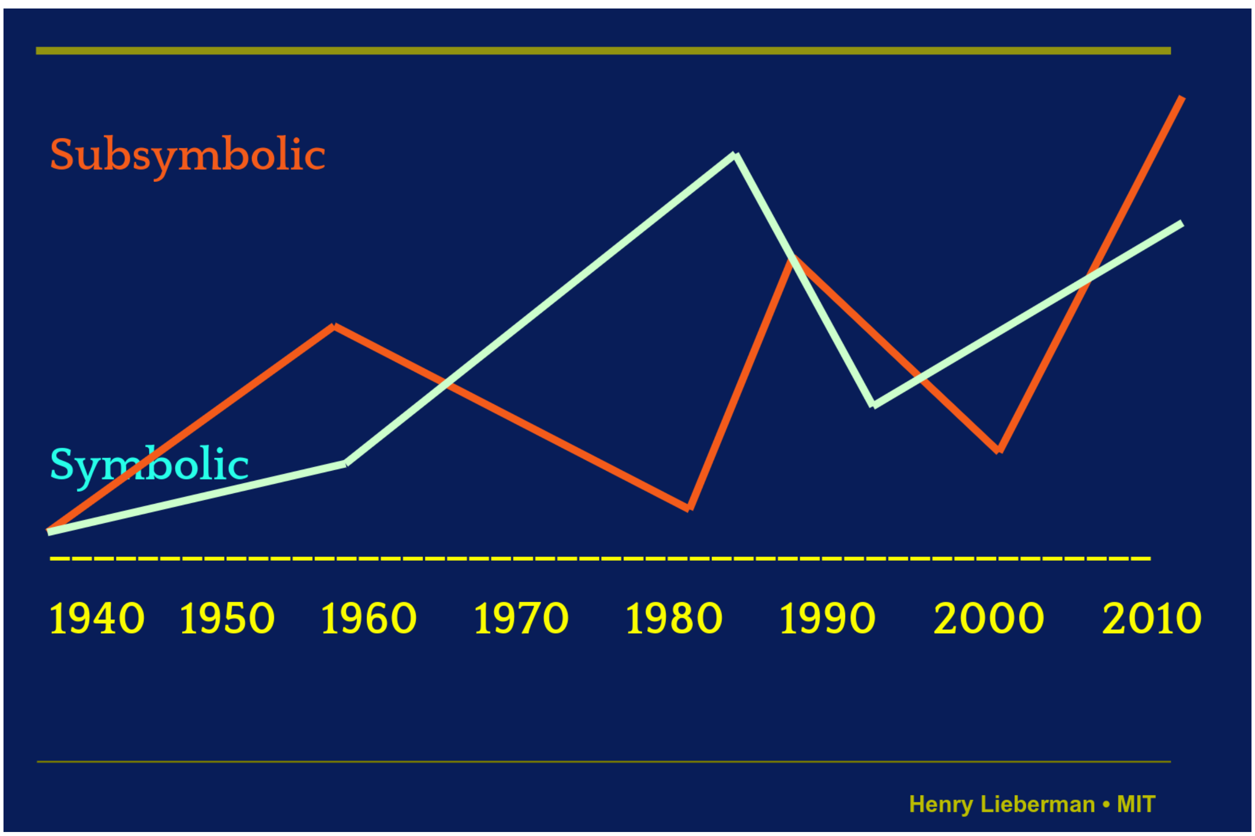
https://en.wiktionary.org/wiki/subsymbolic

This is the entirely 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 buy 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

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 it 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, 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 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 communicate using project coordination tools.

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 (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.

And this may be related to this invisible background history of post-rome western though 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 Khanamn (and then deceased Amos Tversky, the book is about the work they did together but written after Tversky died), where Dr. Kahnamn 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. Shierer, berlin dailies, the rise and fall, Timothy Snider, black earth, ?Anne Applebalm 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 he 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 is 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/Clarke%27s_three_laws
https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
https://en.wikipedia.org/wiki/Isaac_Asimov
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/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/AlexNet
https://en.wikipedia.org/wiki/Dartmouth_workshop
https://en.wikipedia.org/wiki/Frank_Rosenblatt
https://en.wikipedia.org/wiki/John_von_Neumann
https://en.wikipedia.org/wiki/Alan_Turing
https://en.wikipedia.org/wiki/ImageNet#ImageNet_Challenge
https://en.wikipedia.org/wiki/Ronald_Fisher
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/Sally%E2%80%93Anne_test
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://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://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://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://melaniemitchell.me/PapersContent/BurnellEtAlScience2023.pdf
https://melaniemitchell.me/
https://en.wikipedia.org/wiki/ELIZA_effect
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

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 Hofstedter and John Bayes and the inexplicably invisible yet essential Claud Shannon are sometimes missing entirely from a given version of events. (e.g. Claud Shannon was part of the group the 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 jubblies 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 the 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 stupidly, 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 trajically/comically pass each-other in parallel rather than meeting?

The pinball-effect.

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, my 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
- Sigificant 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)
3.3.1 - looking at winograd schemas
3.3.2 - Sally Anne Tasks

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.

Part 4: Goals, Background & Future:

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

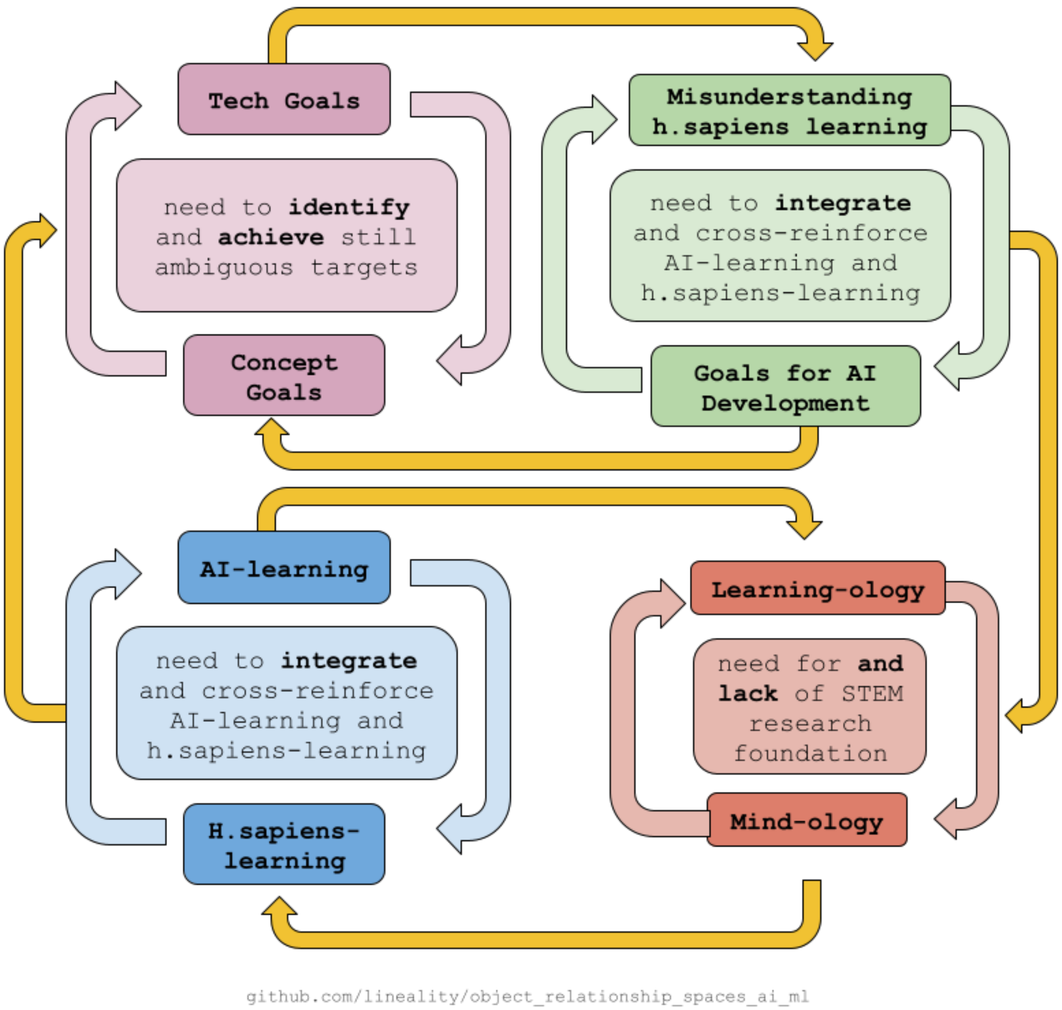
AGI, or Artificial-General-Intelligence, is starting to learn and develop (as of time of writing, April 2023) with its first baby steps coming from "Large Language Models." There are many questions, including one of the first:

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:



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

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
'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

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

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

- 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

4.3 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 knowns) 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 Objet 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 'trolly problem' (regardless of that scenareo'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 frequency 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 aspected 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 Boyscout 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 transfering
- 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 Keti 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 neigher 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 wont 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 Big Blue people assumed it would take a GPT4 type AI with world knowledge and common sense to be able to play chess well, yet Big 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 big 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, he 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=iV1mqabO0bc

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

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 directly 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 is 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 understanding about the problem space and the technology.

For example, Fancois 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 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 Hofstedter 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 Hofstedter 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 fruitway 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 Hofsteder 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, Hofsteder 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 Hofstedter'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

...

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 Hofstedter'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 Hofstedter-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 Hofstedter 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=qbIk7-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.

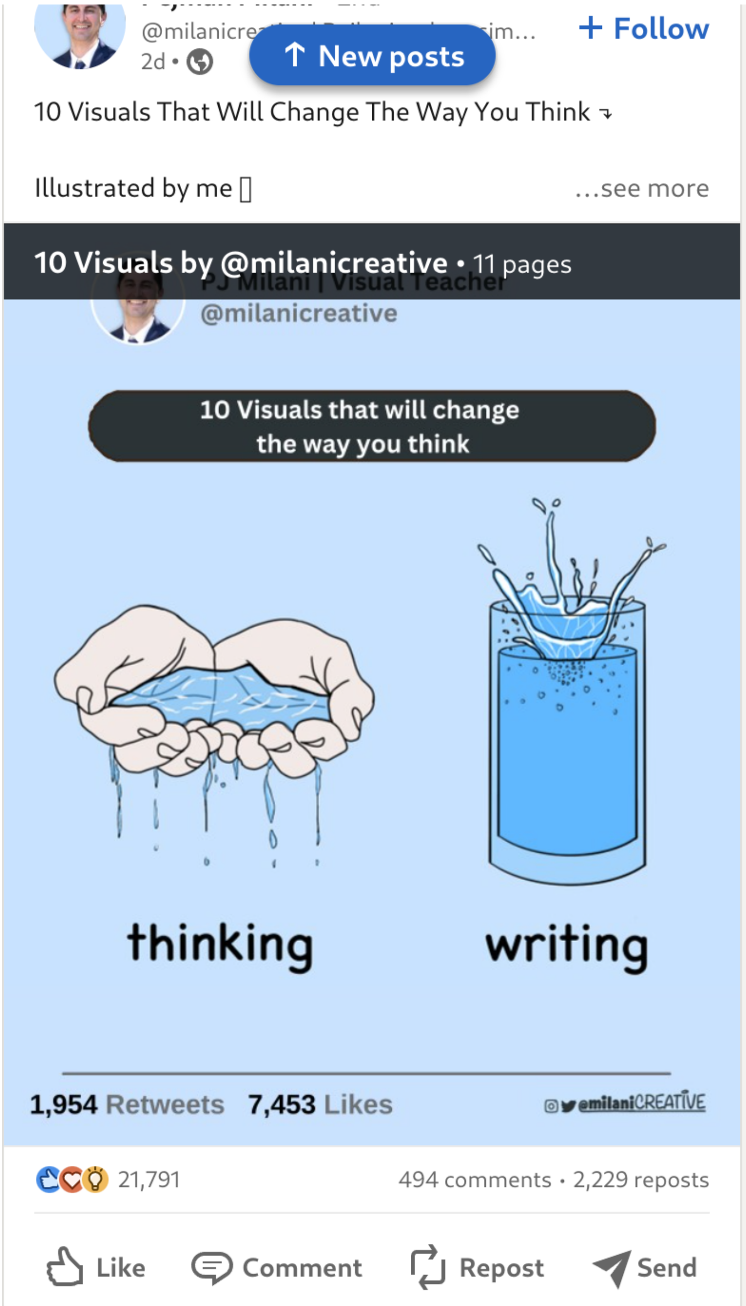
(There may be a rhyme here of the evolution of neural network architectures, where recursive (RNN), then then LSTM models were used to 'retain' threads of learning over time, which then were superseded by 'transformers' (which are the "T" in GPT).)

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 : abbcccdddd:: 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 : abbcccdddd:: 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=qbIk7-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/

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

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.

Appendices

Note: under-construction sections are available in github
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/B07MQX54TW/

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