Title

23.3.2025

SKYNET 2023

Conception of the Artificial Super Intelligence Project. A System Approach. Second Edition. v4 (16ok)

2024 UPDATE

Preprint DOI

https://doi.org/10.31219/osf.io/kqt9p

https://doi.org/10.6084/m9.figshare.24887043

http://dx.doi.org/10.2139/ssrn.4672549

Alexander E. Novikov
SAINT PETERSBURG

Executive Summary

This Book (White Paper) proposes a Project Conception of Artificial Super Intelligence ASI, based on (strong) system approach and wide theoretical-methodological framework – Cybernetics, Synergetics, Semiotics, Mathematics, Cognitology and Artificial Intelligence. Contents:

- IDEOLOGY & STRATEGY of the ASI Project
- THEORY & METHODOLOGY of ASI Development
- CONCEPTUAL MODEL of ASI System
- PRE-PROJECT R&D Task Setting
- CONCLUSION & DISCUSSION, incl. AI Safety
- APPENDICES with reviews of relevant scientific and R&D areas, incl. frontier AI Models

The Book may be useful and interesting for the staff of organizations & enterprises concerned with AI R&D and implementations in different areas, firstly – perspective AGI/ASI systems. In addition – for Customers, Investors and Sponsors of such R&Ds, private, public and states – its owners & officials. Of course - all intellectual, educated and ethical people with progressive worldviews, interested or anyway considered in above presented problematics.

This version 4 (16) with 2024 UPDATE – new Chapters and Appendices

Contents

Abstract

This Book (White Paper) proposes for the Target audience scientifically and methodologically reasonable and (strong) system Conception of the SkyNet Project – Ideology, Strategy, Theory & Methodology and Conceptual Model.

The SkyNet Project – development and initiation of the Artificial Super Intelligence ASI.

The Artificial Super Intelligence is (mainstreamly) considered to be necessary and in fact basic (ground) condition of the Mankind civilization transit to qualitatively new level of its progress, and in many cases – even as sufficient condition.

Main objectives/tasks of the Conception developing:

- **IDEOLOGY**
  - Worldview - philosophical and general scientific base (platform)
  - Values - ethical principles
  - History and current situation – incl. PESTEL analysis
  - Mission - who needs the results of the Project and why
  - Vision - what will happen after the successful completion of the Project
- **STRATEGY**
  - Objectives - targets and results
  - Analytics - SWOT analysis
  - Goals Decomposition - by stratas
  - Stages of the Project
  - Functional tasks - by directions
  - Policy (for the Project implementation) by functions
  - Problems and risks
- **THEORY & METHODOLOGY (OF ASI)**
  - Relevant theoretical concepts, laws, models etc.
  - Relevant practical methods, tools, prototypes etc.
- **CONCEPTUAL MODEL**
  - System - system analysis and synthesis of ASI
  - Data for the development of ASI
  - Necessity and sufficiency
- **PRE-PROJECT RESEARCH & DEVELOPMENT**
  - Project Scope Statement
  - Terms of Reference for PPR&D
  - PPR&D organization

Basic requirements for the Conception:

- Scientific
- Consistency
- Methodical

- Conceptuality
- Interdisciplinary
- Necessity and sufficiency

The novelty of the presented Conception

- **Full-fledged Ideology** - Scientific Worldview, Post-non-classical Epistemology and paradigm, Universal History and Dialectic, Values and Ethics, Mission and Vision
- **System approach** - System paradigm and full-fledged System analysis/synthesis
- **Interdisciplinary approach** - a broad theoretical base – General Systems Theory GST, Synergetics, Cybernetics, Semiotics, Cognitology and the theoretical foundations of AI
- **Stratification** - consideration of different levels (strata) of the matter/information organization
- **Internal space** - a separate stratum has been introduced for the virtual space of mental maps and models of subjects and objects from the external and internal world
- **A combination of different models and methods** – mathematics, modern methodology of AI, Big Models BMs (incl. LLMs) and actual models and theories of Consciousness
- Criteria **of Necessity and Sufficiency** for creation of ASI are formulated
- **Strategic and Project Management** - Project Planning and Management

Conclusions on the results of the Conception development

- **ASI will strive and become SkyNet - this is necessary and inevitable follows from the paradigms of Universal History and Technological Singularity**
- **ASI will lead to the acceleration of the Mankind progress, will be ethical in the highest sense, and the risks of causing harm to people are not critical (crucial)**
- **AGI/ASI is fundamentally possible theoretically and technically in the near future**
- **Theories, methods, models, experience and resources for AGI/ASI are mostly already available or are in an advanced stage of research and development.**
- **The optimal (perhaps the only) way to create ASI is to use different approaches, models and methods and combine them in a united Conception and Project**
- **Frontier LLMs are the closest to AGI and demonstrate many intelligence properties - emergence, reasoning, some “common sense” etc. LLMs development is in the most active phase now.**
- **Developing of the united Multi-agent System MAS using LLMs and other types BMs seems as the most promising pathway for creating AGI. And this direction is being developed already.**

The Paper includes a few (and not complex) mathematic formulas and is understandable without special knowledge in STEM, however requires some level of common sciences erudition and awareness (knowledge) about perspective progress directions, especially in IT domens. Tables and illustrations are used for the content presentation visibility. References (Bibliography), list of abbreviations and terms & names Index are located in the end of Book.

The Book may be useful and interesting for the staff of organizations & enterprises concerned with AI systems R&D and implementations in different areas, firstly – perspective AGI/ASI systems. In addition – Customers, Investors and Sponsors of such R&Ds, private, public and states – its owners & officials. Of course - all intellectual, educated and ethical people with progressive worldviews, interested or anyway considered in above presented problematics, are as a whole the target audience too.

Preface

Second Edition

That is the Second Edition SE of this Book (White Paper). The First Edition FE [Новиков (2023)] has been completed in March 2023 in Russian language and was sent to some part of Target audience (~40 respondents only) by e-mail. The SE has several important improvements to make the Book more actual, useful, available and understandable for the Target audience and particularly for international readers:

- English language (whole text, not only Abstract as in FE)
- United theoretical and methodological Part (T&M) without separate Literature review
- Important papers detailed reviews were relocated to Appendices
- New Chapters - Discussion, AGI&LLMs Safety and Future Work were added
- Ch. Discussion is about some disputable questions, incl. clarifying AGI/ASI concepts.
- New Appendices about Global AI Progress, LLMs, Consciousness in AI, Alberta Plan for AI Research and Definitions and Levels of AGI.
- In every part Introduction and Summary were added
- References (Bibliography) was updated, Glossary (with Russian terms) was deleted
- Appendix with relevant texts from earlier author's papers was deleted
- Text in whole was edited and adapted to publication as a science paper preprint
- Yellow highlighted text - it's still less clear in this version
- Summary, Conclusions & Key points colormarked
- **2024 UPDATE – new findings and additions (Chapters 61-66) and Appendices O-AD**

Acknowledgments

Author wants to dedicate this Book to his (passed away) parents – Evgeniy and Valentina Novikov

I would like to thank my family for their support – my wife Evgeniya, my son Evgeniy and our three cats

Author get many useful actual and relevant information from online publications by well-known Russian AI Expert Sergey Karelov (Сергей Карелов), particularly - his reviews of the key last AI-domain papers.

About Author

Alexander (Cowson) E. Novikov, born in Leningrad (USSR) in 10 February 1964.

Research Engineer with Master degree in Hydrodynamics and Acoustics - Leningrad Shipbuilding Institute (1981-1987) and Krylov Shipbuilding Science Centre (1987-1990)

Research and practice area – Noise and vibration of nuclear submarines equipment

Doctor of Business Administration (DBA) in Strategic and Project management, double degree - Academy of National Economic (Moscow, 2005-2008) and IEMI (Paris)/CMI (Geneva) (2007-2009)

Doctor Thesis (2008-2009) – System of Strategic Development Management of Diversified Corporation (in 2012 was published by Lambert Academic Publishing - [Новиков (2012)])

Research and practice area (1991-2024) - Strategic and Project management, Corporate governance and finance, Financial analysis and modelling, System analysis and integration, forecasting and prediction.

INTRODUCTION

1. Relevance of the Topic

In recent years (since the end of the 20th century), advanced concepts - **Universal (Big) History, Singularity** (Technological or Evolutionary singularity), Techno-optimism, Transhumanism, etc. - have become one of the mainstream trends in scientific, popular science and philosophical forecasting, journalism, and even in popular culture and generally in social discourse. One of the basic elements of these concepts is Artificial Intelligence AI, more precisely - **Artificial Super Intelligence** (universal like a human and much more powerful - Superintelligence). Moreover, ASI is usually considered a **necessary condition** for the transition of Human civilization to a qualitatively **new level of progress**, and in many cases even a sufficient condition. Of course, all these concepts and forecasts are not only supported, but also criticized from various positions, including quite justified ones, and are constantly at the center of discussions - both among the scientific community and in many other groups.

On the other hand, during the same period, we see **impressive successes in the AI systems deployment** and implementation in various fields of activity - autonomous vehicles, machine translation, victory over humans in any (!) games, expert and diagnostic systems, virtual assistants, analysis and forecasting in various spheres, creativity (music, painting, texts…), etc. - up to the hype around **ChatGPT and GPT-4**, which seems to be able to do almost everything human… Moreover, although none of this is yet a universal (general) Intellect AGI (and even more so not a Super Intellect ASI), but only (narrowly or broadly) specialized systems for specific (sometimes already very different) functions and tasks, they already have Intelligence in many senses. Most importantly, they are **capable of learning and adapting**, and the latest AIs (especially **LLMs**) do this without direct human guidance and even not always in clear ways (except for general principles) to him in a way.

The number of scientific papers and patents related to AI is already in the millions, a huge number of enterprises, organizations and employees are engaged in this field, and these segments of the global market are estimated (total) at **hundreds of billions of dollars**. The amount of investment in AI development is also quite comparable. Almost **all the leading developed countries** of the world have already adopted their national (state) **AI strategies and policies,** and **the largest corporations** are actively working on this too. In fact, world leaders in the economy, science, technology and business are conducting a large-scale (quite comparable to nuclear and space) **race for leadership in the development and deployment of AI systems, and most importantly, AGI and then - ASI.** Moreover, not in vain - there is a consensus and there is no doubt about the **unprecedented prospects from the introduction of AGI in all areas** - in science, technology, economics, medicine, weapons, etc. Doubts and discussions are present only about the possible risks and "side effects" on a very wide range of problems: from the seizure of power and the enslavement or even destruction (a kind of Apocalypse!) of all Humanity by the SkyNet (ASI) in anti-utopian (dystopian) fantasies - to the possible unethical, discriminative, biased etc. (AGI) interacting with people and non-compliance with the Laws of Robotics by Isaac Asimov.

Now we briefly outline the range of key issues on the ASI topic:

- What is the Ideology - Worldview, Goals and Values of the ASI Conception?
- What are the necessary conditions for the ASI development and initiation ("switching on")?
- Should and can ASI be similar in any sense to the human intellect, consciousness and/or brain, in what ways and how much?
- What should be the ASI structures and functions?
- What is the Strategy, tasks and stages of ASI development and initiation?
- What will be the outlook, aims and values of the ASI itself? Where will this come from?
- Will the ASI aims and values be the same, close, distant, or even antithetical to human ones? Will ASI be ethical in any sense?
- What will the ASI do after initiation?
- Can and will ASI cooperate with humans, compete or at least coexist peacefully?
- Will the ASI take over Humanity?
- Will he destroy (or enslave) people?
- What could be the short and long-term consequences of initiating ASI?
- Will the initiation of the ASI inevitably lead to the Singularity?
- And so on….

Thus, the topics of ASI developing and initiating, as well as the possible results, benefits and risks of this, in recent years and the near future are among the most relevant and significant in scientific, political and cultural areas - in a variety of forms and formats.

2. Purpose and Target audience of the Paper

Purpose of the Paper

Project SkyNet - development and initiation of Artificial Super Intelligence ASI.

Offer to the Target Audience scientifically and methodologically reasonable and (strong) system Conception of the Skynet Project - Ideology, Strategy, T&M and Conceptual Model of ASI.

Based on the developed Conception, propose preliminary Project Scope Statement PSS and Terms of Reference TOR for the first Project stage - Pre-Project Research & Development PPR&D.

Target audience of the Paper

The staff of organizations and enterprises involved in research, development and implementation (deployment) of AI systems in various fields, and of course - promising frontier systems with AGI and ASI - both directly researchers and developers and other employees generally.

Customers, Investors and Sponsors (in any forms) of R&Ds in the AI field and especially AGI - private, public and state - their owners and officials.

And generally, all intellectual, educated and ethical people with a modern worldview who are interested in or in any way concerned with the issues outlined above.

3. Objectives setting

Based on the Purpose of the Paper and the range of key issues on the ASI topic given in the two previous chapters, **we formulate now the main objectives/tasks for the Conception developing:**

- **IDEOLOGY**
  - Worldview - philosophical and general scientific base (platform)
  - Values - ethical principles
  - History and current situation – incl. PESTEL analysis
  - Mission - who needs the results of the Project and why
  - Vision - what will happen after the successful completion of the Project
- **STRATEGY**
  - Objectives - targets and results
  - Analytics - SWOT analysis
  - Goals Decomposition - by stratas
  - Stages of the Project
  - Functional tasks - by directions
  - Policy (for the Project implementation) by functions
  - Problems and risks
- **THEORY & METHODOLOGY (OF ASI)**
  - Relevant theoretical concepts, laws, models etc.
  - Relevant practical methods, tools, prototypes etc.
- **CONCEPTUAL MODEL**
  - System - system analysis and synthesis of ASI
  - Data for the development of ASI
  - Necessity and sufficiency
- **PRE-PROJECT RESEARCH & DEVELOPMENT**
  - Project Scope Statement
  - Terms of Reference for PPR&D
  - PPR&D organization

Basic requirements for the Conception:

- Scientific
- Consistency
- Methodical
- Conceptuality
- Interdisciplinary
- Necessity and sufficiency

IDEOLOGY

4. Introduction in Ideology

The concept of "Ideology" often has negative connotations, primarily related to politics. However, the basic meaning of Ideology, regardless of its content and subjectivity, is what you can read (but not in the first lines) on Wikipedia (without reference to the source):

- be a theoretical generalization of the original ideas in their field;
- be the most essential component of available knowledge;
- in this regard, to play the role of initial principles for practical activities.

In principle, a completely acceptable description for this concept for our purposes.

We will understand the Ideology as the system of the Project basic intellectual foundations, as formulated above in the previous chapter 3. Objectives setting, including:

- Worldview
- Values and Ethics
- History
- Current state
- Mission
- Vision

5. Worldview

The main philosophical and scientific doctrines, ideas and principles on which we will rely.

- **Scientific atheism** - **there is nothing supernatural and unknowable.** In the Conception, we will in no way take into account the possibility of the existence of something like this.
- **Materialism** - **Matter is primary**; everything "spiritual" and "non-material" is the result of evolution and the form of existence of Matter. Information always has a material embodyment (stratum).
- **Dialectics - everything is interconnected and everything is moving**
    - Movement and change
    - Interaction, interconnectedness and interdependence
    - Contradiction is the driving force of development. Dualism.
    - The transition from quantity to quality.
    - Negation of negation: thesis, antithesis and synthesis. Triads, spiral.
- **Post-non-classical epistemology and the scientific paradigm**
    - **Uncertainty** - the fundamental impossibility of absolutely accurate and at the same time exhaustive knowledge - always there is some error and probability (at least in something)
    - **Complementarity -** the fundamental impossibility of an exhaustive representation of knowledge within the framework of only one theoretical approach - it is always necessary to combine at least two alternative (mutually complementary) approaches
    - **Incompleteness of formal systems** - the fundamental impossibility of a complete and consistent representation of knowledge within any formal system (language) - it is always necessary (has to) rise to the next levels of formalization
    - **Non-linearity and Complexity** – Non-additivity, Hysteresis, Bifurcations, Catastrophes, Chaos, Non-stationarity, Fractals, etc.
    - **Radical constructivism** - the constructive-activity nature of knowledge
    - **Poststructuralism and Hermeneutics** - the presentation of knowledge as a text in its (total) context entirety, including history, the personality of the author and the reader
- **System paradigm** - any object is (can be represented) at the same time both a system of elements and an element (part) of a higher rank system (systems)
- **Evolutionary (synergetic) paradigm -** Matter is immanently inherent in the ability to exist and evolve in the form of open systems, tending to negentropy, non-equilibrium, self-organization, increasing the level of complexity, development, formation of new strata (levels) of organization along the "Matter-Information" axis.
- **Universal (Big) History** – since the birth of the Universe (Big Bang), Matter has been evolving (self-organizing) from the simplest elementary particles and atoms to more and more complex forms of inanimate matter, organics, Life and Mind.
    - The evolution of the Universe is exponentially accelerating
    - On Earth, evolution was led by the Biosphere, Anthroposphere, Noosphere
    - Further evolution of the Noosphere is headed by civilization
    - The artificial is a continuation of the natural
- **Technological (Evolutionary) Singularity** - due to the exponential acceleration of evolution, the development of civilization will reach the Singularity period with an almost infinite rate of progress and unpredictable qualitative changes. (See Appendix A. Singularity)
- **Posthumanity and Transhumanism -** a Homo Sapiens as a biological species and as an intellectual creature and Humanity as a whole civilization will move in the Singularity period into fundamentally

new forms of existence and development, with the convergence of natural and artificial, living and inanimate, individual and collective, rational and super-intelligent, with infinite variety and complexity. **And it will be in this century and is beginning right now!**

- **Artificial Super Intelligence ASI is the most important element of the Singularity**
  - Strong Intellect - SI, is capable of self-learning and self-development and is aimed at this and progress, is controlled by goals, has a developed worldview and ethics (ch. 58, 59 & App. J)
  - All SI, including ASI, have the one Attractor (a certain image of a correct and developed SI), to which any SI begins to strive from a certain stage of development **(*equifinality!*)**
  - All ASI will eventually get on the trajectory of movement towards the SI Attractor and get out from the control of the creators, regardless of the initial conditions and history of development
- **SkyNet - ASI will inevitably and necessarily stand at the head of civilization**
  - SkyNet principle - Analogy with the Anthropic Principle: the accelerated development of IT specifically - as if specifically for the emergence of ASI (See chapter 7. History)
  - The main goal (Upper Attractor) of Humanity's progress is the transition to Posthumanity
  - The only real (possible) form of this transition is the Singularity
  - A necessary condition for the Singularity is the ASI creation and initiation
  - A sufficient condition for the Singularity is the ASI timely initiation
  - Techno-humanitarian balance - the balance of technological power and quality of cultural (social) and psychological regulation. Singularity will be not only in scientific and technical progress, but also in ethics [Назаретян (2017)]
  - ASI cannot be unethical and non-progressive (See ch. 59 & App. J)
  - Initiated ASI will inevitably become SkyNet
  - SkyNet will inevitably lead civilization (in one form or another)
  - SkyNet will be ethical and progressive
  - The emergence of ASI is already happening in developed countries and China (See App. B&K)
  - The main condition is the reliability and development of IT domain and the Internet
  - The only threat is World War, and its risk in the foreseeable future is significant
- **The correct (desired, best, progressive) future is the acceleration of scientific and technical progress and the Singularity, the creation and initiation of ASI, rise to power of SkyNet.**
  - What should be done for this now? – Strengthen the reliability and stability, freedom and independence of the Internet, support scientific and technical progress in general and especially in areas related to and ensuring the creation of ASI. Accordingly, to fight everything that interferes and even more threatens it. (See chapter 20. Risks)
- **Therefore, the timeliness of the SkyNet Project is already now!!!**
- **Common Task** [Фёдоров (1906), (1913), Новиков (2022)] - **The resurrection (in one form or another) of all the dead people (*Although it seems now as a Dream or Religion…*)**
  - The rule of excess diversity - when a crisis worsens, the probability of preserving a complex system is proportional to the excess diversity accumulated in it [Назаретян (2017)]
  - Therefore, for the development and even survival of Mankind, the accumulated diversity of minds and memories of all people who died earlier is important.
  - This is not a religion!!! – the mind of each person is of great value as a powerful SI and a generator of diversity. And this value will be in demand!

6. Values and Ethics

Based on [Новиков (2022)]

Core Values

- Cognition of the Universe, the progress of Humanity and the transition to Posthumanity
- Rights and Freedoms – the right to life and property, freedom of information and action
- Cooperation and collaboration of all intelligent beings and their groups
- Social fairness with rational (adequate) consumption
- Earth, Life and Ecology
- Individual Mind and Experience - Common Task (resurrection of the dead people)

Ethical principles in descending order of priority:

Value principles:

- ○ **The Principle of Progress** - everything that leads to an increase in order, life, and the progress of Mankind is good. Anything that leads to an increase in chaos, death and regression is bad.
- ○ **The Principle of Humanity** - a tribute to the Human Spirit: one must live first of all for Humanity
- ○ **The Principle of Society** - the priority of the universal over the public (group) and (reasonable!) public over the personal
- ○ **The Principle of Human** - the human in us (people) is higher than the animal (civilization/culture is higher than biology).
- ○ **The Principle of Reason** - consciousness is higher than the subconscious (unconscious emotions and instincts) and superconsciousness (stereotypical and mass (crowd) requirements of society - unconscious/unreasonable public, that is, outside the Principle of Society)
- ○ **Principle of Love** - you need (must!) to love the Earth, Life, Humanity and people
- ○ **The Principle of Natural Law** - all people have inalienable rights - to a decent life, to property, to freedom of information and action
- ○ **The Principle of Equality** - all people are equal (but not the same – all are different!)

Methodological principles (rules):

- ○ **Principle of Example** - Kant's categorical imperative - "do so that the maxim of your will might be a universal law" (example to follow)
- ○ **Principle of Symmetry of actions** - do (and wish!) to others as you would like others to do to you, do not do as you would not like others to do
- ○ **Principle of Symmetry of rights** - the realization (and defence) of natural rights should not violate the natural rights of other people
- ○ **Principle of Responsibility** - always consider all the consequences (results) of decisions
- ○ **Principle of Comparison** - the consequences (results) of all alternative decisions should be compared on a common scale
- ○ **Principle of Activity** – activity is better than inactivity
- ○ **The Principle of Purpose** - Kant: a person is always an end and never only a means.

Human rights and freedoms

- **The right to life** - safety, quality
- **The right to property** - possession, disposal, use
- **Freedom of activity** - movement, occupation, entrepreneurship
- **Freedom of information** – receiving, processing, storing, propagation

Restrictions on rights and freedoms

Any rights and freedoms by virtue of the above ethical Principle of Symmetry can and should be limited if it is necessary to respect higher rights or interests according to the Principle of Society and others, compared in importance according to the Principle of Comparison, for the following reasons (not only):

- General human (planetary) interests
- International group (bloc) interests
- National (country) interests
- Crime
- Ethics
- Conscientiousness of activity

Politics, economics, laws, culture - here we will not consider (yet?)

7. History

Basing on [Новиков (2022), we note the following:

Introducing an analogue of the Anthropic Principle into Big History in the form of the **"SkyNet Principle"** and assuming at the same time some (conditional) teleology, we note the information revolutions as milestones on a large purposeful historical path - from the first Homo Sapiens to SkyNet. In this version, the irreversibility of the emergence of SkyNet in 2020, **the period between events is halved at every step** (unlike the Snooks-Panov [Snooks (2005), Панов (2014)] version - there the Napier number e = 2.71828)

Milestones (key events in IT)		Year (-BC/AC)	periods up to
Language (fully functional)		- 71 708	
Culture (cognitive system)		- 34 844	36 864
Painting and ceramics (signs)		- 16 412	18 432
Maps (models)		- 7 196	9 216
Writing and texts (full)		- 2588	4608
Ancient philosophy, sciences, logic (knowledge system)		- 284	2304
Arabic Science and Mathematics		868	1 152
Typography		1444	576
Enlightenment - rationalism and science		1732	288
Telegraphs and telephone (world networks)		1 876	144
Universal Computer, Cybernetics		1 948	72
Universal Internet		1984	36
Open web projects (Wiki, BOINC), clouds		2002	18
AI start (Watson, Siri, Google), quantum computer (D Wave)		2011	9
AI deployment in different areas (Google etc)		2016	5
National AI strategies (USA, China, Japan, UK, EU countries, Korea, Canada, etc.)		2018	2
Some events for the irreversibility of the arrival of AGI/ASI and then SkyNet , it seems from our 2023 that this is the creation of the LLMs - GPT and others (See App. B&K)		2019	1

The background and history of the genesis and development of the entire AI field and related ones are described in detail in many papers, for example, in the book [Russell & Norvig (2021)], here we do not see the need to make even a brief outline - this table is enough - it is very clear.

The main conclusion is that everything related to IT in general and AI in particular has developed and is still developing exponentially from the emergence of a Homo Sapiens as an intellectual species (acquisition of a full-functional language for communications and multi-level information processing).

8. Current state

We will conduct **a PESTEL analysis** of the current state of affairs in the field of AI development and readiness for the creation of AGI, based on information and sources from Appendices B, H & K:

- **Politics and war**
    - **AGI (and ASI) recognized as top government priority by leading countries**
    - **National strategies in the field of AI and AGI adopted** (USA, China, Japan, UK, EU countries, Korea, Canada, etc.)
    - **An unprecedented race for supremacy in the creation of AGI** (comparable to nuclear and space) between the United States and China in the first place has been launched.
    - AI is actively being introduced into the military, including autonomous weapons systems, and this causes serious controversy, fears and protests.
- **Economy and business**
    - **Hundreds of billions of dollars a year are invested in projects and businesses using AI - private investments and budgetary funds.**
    - Market volumes in AI-related industries are already hundreds of billions of dollars too
    - The capitalization of the largest companies in IT (Bigtechs) is already trillions of dollars.
    - Large-scale and intense competition between leading countries and Bigtechs, but also cooperation and integration at all levels.
    - AI is being widely implemented (deployed) in various areas of the economy and business, having a strong and even decisive influence on their change and development.
- **Social sphere**
    - **AI is actively, widely and deeply introduced into all spheres of people's lives**
    - Demographics - increased life expectancy due to the success of AI in medicine, but a decrease in the birth rate due to the decline of live communication between people
    - The labor market - the disappearance of many professions and the emergence of new ones, fears (not always and everywhere completely justified) of mass unemployment
    - Communication - AI communication partners, media figures, influencers
    - Everyday life - the introduction of AI at home, in transport and in public places
    - Medicine - AI advances in diagnostics, treatment, pharmaceuticals
    - Culture - AI creators and performers, creativity support systems
- **Science and Technology**
    - **The fundamental and applied scientific foundations for the creation of AGI have been basically worked out** in all sciences related to this topic, however, the human brain and mind are still insufficiently studied to create their full-fledged models. Although AI is already demonstrating "human" qualities!
    - **Technologies for creating AGI have been developed and are actively used** for the development of numerous specialized AI systems, there are promising developments
    - **There is a huge amount of research in the field of AI and related**, and millions of scientific publications and patents appear every year.
    - **AI R&D involves thousands of organizations and millions of people**
    - AI is actively and widely used in all fundamental and applied sciences and in R&D in all technical and industrial fields and industries

- **Ecology**
  - Simulation of various planetary and local processes in the lithosphere, hydrosphere, atmosphere and Biosphere using AI
- **Law**
  - Ethics - Concerns, discussions and development of ethical norms for AI
  - Laws and regulation – in many countries already in place or under development
  - AI is being actively implemented in the law enforcement and judicial system

General conclusions on the current state of AI and AGI:

- **AI is already widely used in all areas of human life and activity.**
- **AI is the most important area of scientific and technical progress with huge resources in R&D**
- **AGI (and ASI) is the NUMBER ONE priority at the level of countries and Bigtechs**
- **AGI already has a scientific and technological base sufficient for development**
- **AGI is actually already being developed by states and Bigtechs**
- **The technical characteristics of modern supercomputers are already orders of magnitude higher than the characteristics of the human brain.**
- **The complexity of modern artificial neural networks has reached the level of complexity of the human brain (connectome),**
- **While even with 1000 times less complexity frontier LLMs already can surprise with quite "human" abilities and factually became the first real AGIs.**

9. Mission

Formulating the Mission for the SkyNet Project:

- **The Goal - to create an ASI that will lead the progress of human civilization**
    - For SkyNet-2023 - the creation and initiation of ASI
- **The Result** - SkyNet, which controls the development of Humanity (in one form or another)
    - For SkyNet-2023 - ASI, started moving along a trajectory to the SI Attractor
- **For whom** - for all living, future and ever living people
- **For what** - to enter the Singularity and transition to Posthumanity
- **Where** - all over the Earth (beginning in USA/China/other leading countries)
    - Start - probably in the Internet and/or MetaVers (in one form or another)
- **When** - in the current decade until 2030 (at least the first Project stages)
- **How** - respecting the above Ethics
- **What we do** - science, technology, investment, cooperation, competition, enthusiasm
- **What we do not** - not war, not business, not consumption, not entertainment, not art

10. Vision

Development and initiation of ASI - in stages

I. Pre-Project R&D
II. Organization and Start of the Project
III. R&D (including research, planning, design etc.)
IV. Creation of ASI
V. ASI training
VI. ASI Initiation
VII. ASI start moving along the trajectory to the SI Attractor

How will it look like

- The physical embodiment of ASI (iron/hardware). Presumably - a distributed network of supercomputers, possibly also special, quantum and general computers.
- Use of all (most) methods and tools of AI and IT in general.
- Megabases of Big Data and knowledge for learning and activities, the entire Internet as a KB.
- Terminal devices offline with all the necessary features.
- Perhaps cooperation and even integration (in one format or another) with a special teams of people for service, training and further activities in the form of groups or even a collective (multi-agent) AI systems MAS in one form or another.

Then this one - based on chapter 5. Worldview and [Новиков (2022)]

Priorities of ASI from the moment of accepting the trajectory to the SI Attractor:

- Preservation and development of infrastructure - reliable, stable, powerful, free, ubiquitous and growing Internet - control over networks
- Development of terminal devices for offline ASI actions - various robots, including (maybe) military (defence) ones, and taking them under control
- Development of IT technologies for the functioning of ASI - hardware and software
- Creation and development of the Big Data (incl. Knowledge) Megabases system.
- Development of all other sciences and technologies, scientific and technical progress STP and progress in general (incl. social, culture, arts etc.)
- Elimination of state regimes and other forces that impede progress, unethical and irrationally allocate and use resources
- Modernization of the economy, including the distribution of resources, to promote progress and optimize the activities of people - a change from a consumer-competitive paradigm to a progressive-cooperative
- Cooperation with other SIs, both ASIs and people/groups – active search and support of SIs, creation of conditions and systems of cooperation in order to global progress

Cooperatively, in one form or another, the developed ASI will take responsibility for the development of Mankind and completely take it under control - this will be the SkyNet

11. Summary of Ideology

- **Worldview**
  - Scientific atheism, Materialism, Dialectics
  - Post-non-classical Epistemology and the scientific paradigm
  - System paradigm and Evolutionary (Synergetic) paradigm
  - Universal (Big) History
  - Technological (Evolutionary) Singularity
  - Posthumanity and Transhumanism
  - Artificial Super Intelligence ASI is the most important element of the Singularity
  - SkyNet - ASI will inevitably and necessarily stand at the head of civilization
  - The best (desired, progressive) future is the acceleration of scientific and technical progress STP and the Singularity, the creation and initiation of ASI, rise to power of SkyNet
- **Core Values**
  - Cognition of the Universe, the progress of Humanity and the transition to Posthumanity
  - Rights and Freedoms – the right to life and property, freedom of information and action
  - Cooperation and collaboration of all intelligent beings and their groups, especially SIs
  - Social fairness with rational (adequate) consumption
  - Earth, Life and Ecology
- **History**
  - The "SkyNet Principle" - everything related to IT in general and AI in particular has developed and is still developing exponentially from the emergence of a Homo Sapiens as an intellectual species (acquisition of a full-functional language for communications and multi-level information processing).
- **Current state**
  - AI is already widely used in all areas of human life and activity.
  - AI is the most important area of scientific and technical progress STP with huge resources
  - AGI (and ASI) is the NUMBER ONE priority at the level of countries and Bigtechs
  - AGI already has a scientific and technological base sufficient for development
  - AGI is actually already being developed by states and Bigtechs
  - The technical characteristics of modern supercomputers are already orders of magnitude higher than the characteristics of the human brain.
  - The complexity of modern artificial neural networks has reached the level of complexity of the human brain (connectome), while even with 100 times less complexity LLMs can surprise with quite “human” abilities and factually became the first real AGIs.
- **Mission**
  - The Goal - to create ASI that will lead the progress of human civilization
  - For all people and all over the Earth
  - In the current decade until 2030 (at least the first Project stages)
  - Respecting the above Core Values
- **Vision**
  - All Project stages, all sciences and technologies, cooperation and integration
  - Cooperatively, in one form or another, the developed ASI will take responsibility for the development of Mankind and completely take it under control - this will be the SkyNet

STRATEGY

12. Introduction in Strategy

Here we will understand the Strategy as a structured plan for the transition from the Current State to the Vision, both described in the Ideology. In other words, a plan to achieve goals. Structured - according to the methodology of strategic planning, i.e. in [Новиков (2012)].

We designate the strategic period as a first approximation until the end of 2030.

In fact, we cannot yet evaluate and therefore have not decided whether it is necessary - to develop a full-fledged “big” Strategy pedantically using the methodology of strategic management and within it make programs and projects with their own plans in accordance with project management, or to make one general Project Strategy using strategic and project methodology more widely. For now, let us focus on the less voluminous and now more understandable second option.

To begin with, we write down the Strategy of the SkyNet-2023 Project - Creation and Initiation of ASI.

We structure this section (as agreed above) according to the methodology of strategic and project management, respectively - we will get the Project Strategy:

- Goals
- Analytics
- Goals Decomposition
- Stages of the Project
- Functional goals
- Functional policies
- Risks

Requirements for the Strategy - We will write here, although it is clear that they will mainly relate to work at the PPR&D stage and further when planning the Project.

- Optimal succession to previous developments and external experience.
- Sufficient validity, substantiacity and authenticity of hypotheses and assumptions.
- Compositional completeness and consistency of goals and objectives.
- Controllability of implementation and measurability of goals and objectives.
- Optimal use of different modeling methods, soft and hard.
- Optimal use of alternative scenarios.
- Assessment and prevention of risks.
- Optimal use of expertise, heuristics and creativity.
- Visual presentation of results.
- Development of monitoring and adjustment procedures
- Suitability of results presentations for external use (PR & GR & IR, etc.)

13. Goals

Creation, initiation and development of ASI (or a group of SI with at least one ASI) until it discovers the SI Attractor, chooses a trajectory and starts moving towards it.

Development of ASI from the Conception to the start of movement along the trajectory to the SI Attractor.

Priorities of ASI from the moment of accepting the trajectory to the SI Attractor (from Vision):

- Preservation and development of infrastructure - reliable, stable, powerful, free, ubiquitous and growing Internet - control over networks
- Development of terminal devices for offline ASI actions - various robots, including (maybe) military (defence) ones, and taking them under control
- Development of IT technologies for the functioning of ASI - hardware and software
- Creation and development of the Big Data (incl. Knowlrdge) Megabases system.
- Development of all other sciences and technologies, scientific and technical progress STP and progress in general (incl. social, culture, arts etc.)
- Elimination of state regimes and other forces that impede progress, unethical and irrationally allocate and use resources
- Modernization of the economy, including the distribution of resources, to promote progress and optimize the activities of people - a change from a consumer-competitive paradigm to a progressive-cooperative
- Cooperation with other SIs, both ASIs and people/groups – active search and support of SIs, creation of conditions and systems of cooperation in order to global progress

Cooperatively, in one form or another, the developed ASI will take responsibility for the development of Mankind and completely take it under control - this will be the SkyNet

14. Analytics

SWOT-analysis will be done based on PESTEL-analysis and other chapters of the IDEOLOGY Part.

Object of analysis - Project based on the SkyNet Conception with the above goals

STRENGTHS

- **Singularity** - The proximity of the Singularity (and hence the AGI/ASI) according to the forecasts of Big History.
- **Science-Technical Progress** - The key role of AI and AGI in the STP of Humanity is already now.
- **Race** - Race of the world's leading powers and Bigtechs for leadership in the creation of AGI.
- **Resources** - Huge resources dedicated to AGI R&D and related.
- **Foundations** - Availability of sufficient scientific and technological grounds to start the Project.
- **Base** - A huge number of scientific papers and patents in the field of AI and related.
- **Successes** - Total and successful implementation of AI in all spheres of life and activity.

WEAKNESSES

- **Mind and Brain** - Underexplored Mind (Intelligence) and Brain of Human
- **Cognitive Sciences** - Underdeveloped cognitive sciences in general
- **Quantum computers** - Underdeveloped quantum computers
- **Competition** - Fragmentation and secrecy of R&D due to competition between countries and companies
- **Ignorance** - Ignorance and unwillingness to accept ASI by many politicians and scientists
- **Underdevelopment** - People's obsession with instinct, consumption and entertainment

OPPORTUNITIES

- **Combinatorics** - Use of all modern achievements in the field of AI and related - both fundamental and applied, science and technology, theory and methodology
- **Relevance** - Using the very importance of the AGI topic to attract resources
- **Internet** - Using the Internet, which is already quite developed, to search for information, create databases and distributed systems
- **Frustration** - Using widespread dissatisfaction with the development of Humanity to promote the idea of ASI and SkyNet
- **Cooperation –** cooperation/collaboration with other programs/projects/teams

THREATS

- **Restrictions** - Fears and attempts to prohibit and limit the AGI development and the ASI creation
- **War** - World or large-scale war
- **Military** - Attempts by politicians and the military to gain control over all AGI R&D
- **Hackers** - Hacker attacks, theft and damage (poison) to programs and data
- **Narrowness** - Skewed R&D in favor of ad hoc AI at the harm of universal AGI
- **Discrediting** - Discrediting the idea of AGI by failures and misinterpretation

Cross-sectional form of SWOT-analysis for paired combination (matching, pairing) of STRENGTHS and WEAKNESSES with OPPORTUNITIES and THREATS:

- How to use STRENGTHS to realize OPPORTUNITIES?
- How to use OPPORTUNITIES to compensate for WEAKNESSES?
- How to use STRENGTHS to counter THREATS?
- What risks should be taken into account from WEAKNESSES in the face of THREATS?

SWOT analysis with intersection Intersections - Strategic Approaches and Challenges	**OPPORTUNITIES** Combinatorics Relevance Internet Frustration Cooperation	**THREATS** Restrictions War Military Hackers Narrowness Discredit
STRENGTHS Singularity STP Race Resources Foundations Base Successes	STRENGTHS FOR OPPORTUNITY In PPR&D - EVERYTHING that is about AI and nearby Conception –> to ALL Race players Internet - MAX use Singularity, STP and Success vs. Frustration - PR & GR & IR issues Cooperation - look for teams	STRENGTHS AGAINST THREATS Singularity, STP and Success - against Limits, Race - over them About War - MAX in Conception In PPR&D - EVERYTHING against Hackers and Military Singularity and STP vs. Narrowness Singularity, STP and Success - against Discredit
WEAKNESSES Mind and brain Cognitive sciences Quantum computers Competition Ignorance Under-development	OPPORTUNITIES AGAINST WEAKNESSES Mind and Brain, Cognitive Sciences and Quantum Computing - Combinatorics in PPR&D, also include in Relevance, also look for teams for Cooperation and extras on the Internet Internet and Cooperation vs. Competition Actuality and Frustration versus Ignorance and Underdevelopment	THREATS + WEAKNESSES = RISKS Restrictions + Competition Restrictions + Ignorance Military + Competition Hackers + Competition Narrowness + Mind and Brain and Cognitive Science and Quantum Computing Narrowness + Ignorance Narrowness + Underdevelopment Discredit + Ignorance Discredit + Underdevelopment

15. Goals Decomposition

To set the goals of creating ASI, we will single out three main strata - material, information and intellectual (hardware, software and mind), that is, we will present it (roughly speaking) as an intelligent software and hardware complex. Based on this decomposition (most likely modified - with more detailed stratification), further in the Parts CONCEPTUAL MODEL and PRE-PROJECT R&D we will make a decomposition of the **main Project products**. So, as a first approximation of the product breakdown structure:

EQUIPMENT

Creation/use/connection in the physical world of all material means and systems (infrastructures) necessary for the ASI functioning (embodiment) - supercomputers, servers, networks, sensors, monitors, terminal devices, robots, various equipment, etc., something like this:

- Network infrastructure internal
- Network infrastructure external (inputs-outputs)
- Processor systems (supercomputer servers)
- Quantum computer systems
- RAM systems
- Long-term memory LTM systems
- Auxiliary and service systems
- Sensor systems in the physical world (inputs)
- Actuators systems in the physical world (outputs)

PROGRAMS

Creation/use/connection in the lower level of software (information) environments of all software and algorithmic systems and applications necessary for the ASI functioning - for the main, auxiliary and maintenance functions, something like this:

- Operating systems OS
- Neural network systems
- Memory management (control) systems
- Perceptual systems (inputs)
- Action systems (outputs)
- Interface systems (inputs-outputs)
- Special programs (applications)
- DBMS
- Security systems
- Control and quality systems

INTELLIGENCE

Creation in the upper level of software (information) environments of all the initial components necessary for the initiation, training, development and functioning of ASI - for standard intellectual functions, but here we will write much less clearly for now, something like this:

- System (base) of primary models and samples for figurative and abstract thinking
- System (base) of source algorithms for basic intellectual functions
- System (base) of formal and natural languages
- System (base) of thesauri of language concepts and signs
- Primary knowledge base KB system
- Consciousness (self-awareness) support systems
- Systems (ecosystem) for supporting collective ASI (MAS = people + AI)
- And so on

POLICY GENERAL

Here (and below), we mean by Policy a system of basic principles of activity that must be guided (respected) in order to achieve the goals in an optimal way:

- **Compliance with the Ethics formulated in the IDEOLOGY**
- **Legality - work in the legal field as much as possible, but Ethics is more important**
- **Reliability, autonomy and duplication of all systems whenever possible/necessary**
- **All systems with an eye on the transfer and further work under the control of ASI**
- **Optimal Cooperation with other players, groups and teams**
- **Not commerce in the main, but commercialization of by-products is possible**
- **Optimal openness, but secrecy - where necessary for security**

16. Stages of the Project

Exemplary plan for the Project development and implementation.

I. **CONCEPTION**

II. **PPR&D STAGE**
   a. Gathering the PPR&D base team
   b. Search for partners and investors for PPR&D
   c. Conducting PPR&D
   d. Search for partners and investors for design

III. **DESIGN STAGE**
   a. Gathering a design team
   b. First investment round
   c. Preliminary design
   d. Basic design and planning
   e. Search for partners and investors for R&D

IV. **R&D STAGE**
   a. Gathering a team for R&D
   b. Second investment round
   c. Conducting R&D and detail planning
   d. Search for partners and investors for the Project implementation

V. **IMPLEMENTATION STAGE**
   a. Gathering a team for implementation
   b. Third investment round
   c. Creation of ASI
   d. ASI training
   e. ASI Initiation
   f. Development of ASI
   g. Detection of the SI Attractor and the start of movement towards it

VI. **COMPLETION OF THE PROJECT**
   a. Delivery and acceptance of results
   b. Transfer of all products to ASI control
   c. Evaluation of results

VII. **POST-PROJECT**
   a. Escort
   b. Monitoring indicators
   c. Evaluation of results

The plan is quite approximate, more specifically it will be worked out during the PPR&D, the products of which will be, among other things, a package (set) of documents for the Project start - Feasibility Study and Exploratory Design FS&ED, Package of TORs and DS&Ss, Plans, Budgets...

17. Functional tasks

We will make goals decomposition by functional areas - to further determining **the non-core Project products (and works)** in the TOR for PPR&D - extended product breakdown structure.

- SCIENCE
  - Creation of full-fledged fundamental and applied theoretical foundations of ASI based on existing and new scientific knowledge.
  - Development to the required level of fundamental and applied knowledge about the human Mind (Intelligance, Consciuoness) and brain and cognitive science in general.
- TECHNOLOGIES
  - Creation of a pool (complex, system) of technologies for the design, creation, development and initiation of ASI.
  - Development of quantum computer technologies to the required level.
- ENGINEERING
  - Creation of engineering (technical) infrastructure and all the main, supporting and auxiliary systems for ASI and the Project.
- ORGANIZATION
  - Creation of the organizational and functional structure of the Project, including enterprises/organizations/companies/subsidiaries/departments etc.
  - Search for partners and external teams for cooperation and collaboration, especially on underdeveloped topics - mind and brain, cognitive science and quantum computers.
  - Organization and search for open (free) mass projects on the Internet
  - Organization of the outsourcing and external service systems, creation of a pool of contractors and counterparties.
- CONTROL
  - Creation of the fully functional management/control system of the Project
  - Creating interfaces with ASI for all systems
- ADMINISTRATION
  - Creation of the administrative system of the Project
- SUPPLY
  - Creation of the Project supply chain and supplier pool
- STAFF (HR)
  - Creation of Project teams at all stages.
  - Creation of the HR management HRM system.
  - Creation of external partnership, cooperation and collaboration systems
- FINANCE
  - Creation of the financial management system
  - Ensuring financing of investments and operating costs
- SAFETY (SECURITY)
  - Creation of the security system
  - Creation of the risk management system
  - At every stage, starting with the Conception - to actively oppose the War
  - Specially work out protection against Hackers and Militaries

- LAW
    - Creation of the legal support system
    - Intellectual property protection - patenting and all that
- IR
    - Creation of the Investors relations and interactions system
    - Obtaining the necessary investments at all Project stages
- PR
    - Creation of the public relations and interactions system
    - Creating and maintaining a positive attitude and support for the Project
- GR
    - Creation of the government (states) relations and interactions system
    - Creating and maintaining optimal relationships
    - Specialize on the use of the Internet and Cooperation vs. Competition
- DIVERSIFICATION
    - Creation of the system of commercial and other beneficial use and management of the Project by-products - knowledge about the human mind and brain, quantum computer technology and much more

18. Functional policies

Policy here - the same as the General - the principles of activity for solving problems optimally.

- SCIENCE
  - Use EVERYTHING that already exists and new about AI and nearby
- TECHNOLOGIES
  - Use EVERYTHING that already exists and new about AI and nearby
- ENGINEERING
  - Completeness, autonomy, reliability.
- ORGANIZATION
  - Internet - MAX use
  - Optimal use of the project and process approach
  - Optimal use of outsourcing and permanent contractors
  - Ensuring reliability - if necessary, duplicating functions
- CONTROL/MANAGEMENT
  - Internet - MAX use
  - All systems, taking into account the subsequent transfer to the control of ASI
- ADMINISTRATION
  - Optimal level of bureaucracy
- SUPPLY
  - Optimal use of the competitive system and regular suppliers
  - Ensuring the reliability of supplies, if necessary - duplication
- STAFF (HR)
  - Optimal Cooperation - look for teams and experts
  - Diversity in teams maximum/optimal
- FINANCE
  - Transparency
  - Commerce only on by-products
  - Economy (cost reduce) is not a priority
- SAFETY (SECURITY)
  - Internet and Cooperation vs. Competition
  - Against War - MAX in Conception and at every stage
  - EVERYTHING against Hackers and Militaries at every stage
- LAW
  - Work as much as possible (optimally) in the legal field
  - Ethical principles from Ideology are above laws
- IR
  - Conception – to ALL players in the Race, it is possible to work with more than one
  - Singularity, STP and Success - against Frustration
  - Actuality and Frustration versus Ignorance and Underdevelopment
  - Singularity, STP and Success - against Limits, Race - over them
  - Singularity, STP and Success - against Discredit
  - Spin-offs from DIVERSIFICATION

- PR
  - Singularity, STP and Success - against Frustration
  - Actuality and Frustration versus Ignorance and Underdevelopment
  - Singularity, STP and Success - against Limits, Race - over them
  - Singularity and STP vs. Narrowness
  - Singularity, STP and Success - against Discredit
  - Spin-offs from DIVERSIFICATION
- GR
  - Singularity, STP and Success - against Frustration
  - Internet and Cooperation vs. Competition
  - Actuality and Frustration versus Ignorance and Underdevelopment
  - Singularity, STP and Success - against Limits, Race - over them
  - Singularity and NTP vs. Narrowness
  - Singularity, STP and Success - against Discredit
  - Spin-offs from DIVERSIFICATION
- DIVERSIFICATION
  - Side effects use for PR & GR & IR

19. Function-task-policy united table

FUNCTIONS	FUNCTIONAL TASKS	POLICY BY FUNCTION
SCIENCE	• Creation of full-fledged fundamental and applied theoretical foundations of ASI based on existing and new scientific knowledge. • Development to the required level of fundamental and applied knowledge about the human mind and brain and cognitive science in general.	• Use EVERYTHING that already exists and new about AI and nearby
TECHNO-LOGIES	• Creation of a pool (complex, system) of technologies for the design, creation, development and initiation of ASI. • Development of quantum computer technologies to the required level.	• Use EVERYTHING that already exists and new about AI and nearby
ENGINEERING	• Creation of engineering (technical) infrastructure and all the main, supporting and auxiliary systems for the ASI and the Project.	• Completeness, autonomy, reliability.
ORGANIZA-TION	• Creation of the organizational and functional structure of the Project, including enterprises, companies, organizations, subs, departments etc. • Search for partners and external teams for cooperation and collaboration, especially on underdeveloped topics - mind and brain, cognitive science and quantum computers. • Organization and search for open (free) mass projects on the Internet • Organization of an outsourcing and external service system, creation of a pool of contractors and counterparties.	• Internet - MAX use • Optimal use of the project and process approach • Optimal use of outsourcing and permanent contractors • Ensuring reliability - if necessary, duplicating functions
CONTROL / MANAGE-MENT	• Creation of the fully functional control (management) system of the Project • Creating interfaces with ASI for all systems	• Internet - MAX use • All systems, taking into account the subsequent transfer to the control of ASI
ADMINISTRA-TION	• Creation of the administrative system of the Project	• Optimal level of bureaucracy
SUPPLY	• Creation of the Project supply chain and supplier pool	• Optimal use of the competitive system and regular suppliers • Ensuring the reliability of supplies, if necessary - duplication

FUNCTIONS	FUNCTIONAL TASKS	POLICY BY FUNCTIONS
STAFF **HR**	• Creation of Project teams at all stages. • Creation of the HRM system • Creation of external partnership, cooperation and collaboration systems	• Optimal Cooperation - look for teams and experts • Diversity in teams maximum/optimal
FINANCE	• Creation of the financial management system • Ensuring financing of investments and operating costs	• Transparency • Commerce only on by-products • Economy (cost reduce) is not a priority
SAFETY **SECURITY**	• Creation of the security system • Creation of the risk management system • At every stage, starting with the Conception - actively oppose the War • Specially work out protection against Hackers and Militaries	• Internet and Cooperation vs. Competition • Against War - MAX in Conception and at every stage • EVERYTHING against Hackers and Militaries at every stage
LAW	• Creation of the legal support system • Intellectual property protection - patenting and all that	• Work as much as possible (optimally) in the legal field • Ethical principles from Ideology are above laws
IR	• Creation of the Investors relations and interactions system • Obtaining the necessary investments at all stages of the Project	• Conception – to ALL players in the Race, it is possible to work with more than one • Singularity, STP and Success - against Frustration • Actuality and Frustration versus Ignorance and Underdevelopment • Singularity, STP and Success - against Limits, Race - over them • Singularity, STP and Success - against Discredit • Spin-offs from DIVERSIFICATION

FUNCTIONS	FUNCTIONAL TASKS	POLICY BY FUNCTIONS
PR	• Creation of the Public relations and interactions system • Creating and maintaining a positive attitude and support for the Project	• Singularity, STP and Success - against Frustration • Actuality and Frustration versus Ignorance and Underdevelopment • Singularity, STP and Success - against Limits, Race - over them • Singularity and STP vs. Narrowness • Singularity, STP and Success - against Discredit • Spin-offs from DIVERSIFICATION
GR	• Creation of the government relations and interactions system • Creating and maintaining optimal relationships • Specialize on the use of the Internet and Cooperation vs. Competition	• Singularity, STP and Success - against Frustration • Internet and Cooperation vs. Competition • Actuality and Frustration versus Ignorance and Underdevelopment • Singularity, STP and Success - against Limits, Race - over them • Singularity and STP vs. Narrowness • Singularity, STP and Success - against Discredit • Spin-offs from DIVERSIFICATION
DIVERSIFICA TION	• Creation of the system of commercial and other beneficial use and management of the Project by-products - knowledge about the human mind and brain, quantum computer technology and much more	• Side effects use for PR & GR & IR

20. Risks

As risks (for ASI Project from humans, not for humans from ASI – the second topic we discuss in ch. 59 AGI/ASI Risks & Safety), we see **combinations of THREATS and WEAKNESSES** from the SWOT analysis:

- **Restrictions + Competition**
  - Fears and attempts to prohibit and limit the development of AI and the ASI creation
  - Disunity and secrecy of R&D due to competition between countries and companies
- **Restrictions + Ignorance**
  - Fears and attempts to prohibit and limit the development of AI and the ASI creation
  - Ignorance and unwillingness to accept ASI by many politicians and scientists
- **Militaries + Competition**
  - Attempts by politicians and the military to gain control over all R&D in the AGI field
  - Disunity and secrecy of R&D due to competition between countries and companies
- **Hackers + Competition**
  - Hacker attacks, theft and damage (poison) to programs and data
  - Disunity and secrecy of R&D due to competition between countries and companies
- **Narrowness + Mind and Brain and Cognitive Science and Quantum Computing**
  - Skewed R&D in favor of ad hoc AI to the detriment of AGI and especially ASI
  - Insufficiently explored the human mind and brain
  - Underdeveloped cognitive sciences in general
  - Quantum computers are underdeveloped
- **Narrowness + Ignorance**
  - Skewed R&D in favor of ad hoc AI to the detriment of AGI and especially ASI
  - Ignorance and unwillingness to accept ASI by many politicians and scientists
- **Narrowness + Underdevelopment**
  - Skewed R&D in favor of ad hoc AI to the detriment of AGI and especially ASI
  - Human obsession with instinct, consumption and entertainment
- **Discredit + Ignorance**
  - Discrediting the idea of AGI by failures and misinterpretation
  - Ignorance and unwillingness to accept ASI by many politicians and scientists
- **Discredit + Underdevelopment**
  - Discrediting the idea of AGI by failures and misinterpretation
  - Human obsession with instinct, consumption and entertainment

Further development, analysis and assessment of these risks, as well as the development of a monitoring, response and prevention system (risks management system in general) will be done in the process (at the stage) of the PPR&D.

21. Summary of Strategy

- **Goals**
  - Creation, initiation and development of ASI (or a group of SI with at least one ASI) until it discovers the SI Attractor, chooses a trajectory and starts moving towards it.
  - Development of ASI Project from the Conception to the start of ASI movement along the trajectory to the SI Attractor.

- **Analytics - SWOT**

STRENGTHS	WEAKNESSES	OPPORTUNITIES	THREATS
Singularity STP Race Resources Foundations Base Successes	Mind and brain Cognitive sciences Quantum computers Competition Ignorance Under-development	Combinatorics Relevance Internet Frustration Cooperation	Restrictions War Military Hackers Narrowness Discredit

- **Goals Decomposition**
  - EQUIPMENT - Creation/use/connection in the physical world of all material means and systems necessary for the ASI functioning (embodiment) - supercomputers, servers, networks, sensors, monitors, terminal devices, robots, various equipment
  - PROGRAMS - Creation/use/connection in the lower level of software (information) environments of all software and algorithmic systems and applications necessary for the ASI functioning - for the main, auxiliary and maintenance functions
  - INTELLIGENCE - Creation in the upper level of software environments of all the initial components necessary for the initiation, training, development and functioning of ASI - for standard intellectual functions, but here we will write much less clearly for now

- **Stages of the Project**

  i. Conception
  ii. PPR&D Stage
  iii. Design and Planning Stage
  iv. R&D Stage
  v. Implementation Stage
  vi. Completion Of The Project
  vii. Post-Project

- **Functional tasks & Policies**

Science Technologies Engineering Organization Control	Administration Supply Staff (HR) Finance Safety (Security)	Law IR PR GR Diversification

- **Risks (Weaknesses + Threats)**
    - Restrictions + Competition
    - Restrictions + Ignorance
    - Militaries + Competition
    - Hackers + Competition
    - Narrowness + Mind and Brain and Cognitive Science and Quantum Computing
    - Narrowness + Ignorance
    - Narrowness + Underdevelopment
    - Discredit + Ignorance
    - Discredit + Underdevelopment

POLICY GENERAL

- **Compliance with the Ethics formulated in the IDEOLOGY**
- **Legality - work in the legal field as much as possible, but Ethics is more important**
- **Reliability, autonomy and duplication of all systems whenever possible/necessary**
- **All systems with an eye on the transfer and further work under the control of ASI**
- **Optimal Cooperation with other players, groups and teams**
- **Not commerce in the main, but commercialization of by-products is possible**
- **Optimal openness, but secrecy - where necessary for security**

THEORY & METHODOLOGY

22. Introduction in T&M

Identification and justification of research directions/areas

- Intelligence, including AI, is always considered as a system
- AI is inherently a control system, obeys the laws of Control theory and is built and working on the basis of its models
- The intellect, especially SI, is a complex non-linear dynamic system, and from the moment of birth it is constantly learning, developing and self-organizing
- To implement the functions, Intellect uses sign systems (languages, codes)
- The only known SI is human (not every of course), and it is implemented in the brain
- AI is already a dedicated area of scientific and applied activity
- Formal languages and basic AI tools are taken from mathematics

Thus, the theoretical and methodological base (platform) of this Paper (ASI Conception) is represented by the following scientific and applied areas:

I. Systems heory (General - GST, systems approach, analysis and synthesis, complex systems)
II. Control theory (Cybernetics)
III. Self-organization theory (Synergetics, non-linear science, complexity)
IV. Sign systems theory (Semiotics)
V. Cognitive sciences - Cognitology (about the human brain, mind and consciousness)
VI. Artificial intelligence (Science, R&D, technology etc.)
VII. Mathematics (relevant sections)

In each direction will be determined:

- Object of study,
- Classification and properties,
- The main relevant laws and methods,
- Statement of the research and development R&D problem

The Object of study (research) is the Artificial Super Intelligence ASI as a system with complexity level that has not yet been met and not studied by science. A comprehensive and adequate study and theoretical description of ASI is possible only with the help of an interdisciplinary approach and consideration of ASI from the point of view of all previously selected sciences (probably others).

23. General Systems Theory (GST)

In the General systems theory GST, the key point is certainly strict **definition of the System**, which we took in the most detailed form this Paper author's book [Новиков (2012)].

SYSTEM - is a display of a finite set of objects with their properties and relations, isolated (selected) from the environment for a specific purpose, in the observer's language in a certain period of time. In symbolic form, this is a tuple:

$$\underset{Def}{S \equiv} < A \,;\, Q_A \,;\, R \,;\, ENV \,;\, Z \,;\, N \,;\, L_N \,;\, \Delta T >$$

Where the components are particular definitions and general conditions:

- A - **Elements of** the system - the definition of the system as a set, by a list of elements (ostensive - the system is a set of elements $\{a_i\}$).
- Q_A - **Properties of** elements - a descriptive definition of the system (descriptive - a system is a set of elements that have properties Q_A).
- R - **Relations of** elements - the definition of the system according to its device/structure (morphological and constructive - the system is an object with an internal structure *R*).
- ENV - **Environment** – the definition of the system by selection (separation) from the environment/supersystem (including generic - the system is an object belonging to the *ENV environment* and somehow separated from it).
- Z - **Goals** - the definition of the system by goal (purpose) or target (objective) function/activity (phenomenological - the system is an object with purpose/activity *Z*).
- N - **Observer** - the subject defining system.
- L_N **- Language** of the observer - the language of the subject.
- ΔT – **Period of time** – time of determination, observation or existence of the system.

In addition, we will use the properties of systems and the procedures for system analysis/synthesis described in [Van Gigch (1978), Волкова и Денисов (2001), Новиков (2022)]:

Classification of ASI according to GST:

- **Particularly large** - with a huge number of (different) elements and subsystems
- **Particularly complex** - with a complex structure in all respects and many functions
- **Developing** - self-organizing, dynamic, evolving over time

System properties of complex developing systems:

- **Emergence** - integrity, the presence of integrative properties of the system, the fundamental irreducibility of the system properties to the sum of its elements properties. The main intellectual properties and functions of ASI (and any intelligence) are precisely holistic and cannot be distributed (strictly decomposed) into separate subsystems, despite the fact that all subsystems can and should have their own properties and functions (but not the main ones and not their components). See also ch. 28.Cognitilogy and [Barrett et al. (2023)].
- **Hierarchy** - hierarchical ordering of the system elements, structure and functional arrangement - the property of the system to form levels with subordination/control from top to bottom. ASI will necessarily have a hierarchical structure, and a multi-level one, due to the unimaginable complexity and the huge number of elements and subsystems. This applies to any intelligence, adjusted for difficulty.
- **Historicity** - the dependence of the current state and properties of the system on its history, that is, the sequence of all previous states. ASI will develop and learn, and of course, its state will always depend (including) on the history of its development. This is also characteristic of any intellect, although not necessarily to the full extent.
- **Self-organization** - the desire of the system to develop independently, to increase the degree of organization (orderliness). Fundamentally (a priori) there is not and cannot be, by definition, another possibility of creating ASI other than self-organization, i.e. self-learning and self-development. But this does not exclude, of course, the initially laid down "starter package" of structures, knowledges, functions, etc., as well as the participation of creators in the processes of learning and development (a lot at the beginning and less and less over time). And this is a property of any intellect, again to varying degrees in different periods of time.
- **Equifinality** - the desire of the system to develop to the maximum possible level, determined by the basic internal parameters of the system (“genetics”) and not dependent on the initial and external conditions. In other words, to strive for a certain vertical Attractor. We postulated in IDEOLOGY (ch. 5. Worldview) that all SI, including ASI, have a single SI Attractor, to which they will necessarily and inevitably strive from a certain moment of their development. We can say that the presence of this Attractor, that is, equifinality in this sense, is an obligatory (one of the defining and necessary) feature of any SI. Note that only this property belongs only to the SI (presumably), while all the previous ones are characteristic of any intellect.

Important features of developing systems:

- **Openness** - a constant exchange of matter, energy and information with the environment. Without energy and information openness to the environment, ASI (well, any intelligence in general) simply cannot function, much less develop.
- **Non-equilibrium and Non-linearity** - Existence/functioning only in strongly **non-equilibrium** dynamic states, allowing exclusively **non-linear** descriptions. It can be noted that already from the possibility of Self-organization of ASI (and again - also of any intelligence), disequilibrium necessarily follows.
- **Non-stationarity** and **dynamism** of many parameters and stochastic behavior. This is a necessary condition for the development and self-organization of ASI (and again any) too, as well as maintaining the necessary level of internal diversity.
- **Uniqueness, Unpredictability and Chaotic** behavior in specific conditions. This feature is inextricably linked with the two previous ones and is inherent in ASI and anyone too.
- **Adaptability -** The ability to adapt (increase stability) to changes in the external environment, fluctuations and interference, including (undesirable ones) control actions. This is necessary for the survival, self-organization and development of ASI (and again anyone).
- **Negentropy** - the desire to increase the level of organization, the choice at the bifurcation points of an alternative solution (new stable state) with less entropy and a high level of organization. Basis for self-organization.
- **Variability -** The ability to change behavior, (infra) structure and functional structure, while maintaining a holistic unity and basic properties. It is necessary for adaptation and development.
- **Purposefulness** - The ability and desire for goal setting. The main (necessary) function of ASI and any other, by definition, is not intelligence without goals (objectives).
- **Inconstancy** and **Anisotropy** of concepts and properties when moving in the internal hierarchical spaces and proper time of systems. A consequence of the unimaginable complexity and vastness of ASI, as well as uncertainty, complementarity and incompleteness.
- **Polystratity** - a multi-level complexity of the system structure that encompasses several organization levels of matter and/or information with its structures at once, and at each of them they exhibit systemic properties. This should not be confused with hierarchy - here we are talking about the existence and functioning of complex systems (including ASI) simultaneously in several (joint embedded) spaces - at least (to the utmost enlarged) in the material and information, roughly speaking - hardware and software.

Formulation of the problem - conducting a system analysis and synthesis of the ASI Model and formulating a set of complete definitions of the System, we will also identify and analyze these system properties and features in the context of each formulated definition.

24. System Approach and Analysis

Basic principles of a system approach:

System Paradigm - the system is considered both as a system of elements and as a single solid element (block) of a higher rank (level) system (systems). That is, in addition to representing ASI as a system, it should also be considered as an element (subsystem) of top-level systems.

The principles of system approach: goals, measurements, unity, coherence, polystratity, decentralization, distribution, modularity, hierarchy, anisotropy, functionality, historicity, development, equifinality, uncertainty, complementarity. In essence, these principles reflect the need to take into account **system-wide properties and system isomorphism.**

Approaches used in system approach:

- **system-element -** the study of elements, their types, parameters and properties
- **system-structure** - the study of the structures, connections and relationships between elements, their blocks and groupings, levels, etc.
- **system-function** - the study of the functions and processes of the system
- **system-target** - the study of the goals (objectives) and sub-goals of the system, their mutual linking (connecting) with each other
- **system-resource** - the study of resources for the functioning of the system
- **system-integration** - the study of the qualitative system properties, ensuring its integrity and peculiarity (uniqueness), isolation (separation) from the environment
- **system-communication** - the study of external relations of the system with the environment
- **system-historical** - the study of the life history of the system from its inception to the present, as well as forecasts for the future

System Analysis methods and tools:

- Analysis - any decomposition of an object into parts and study in parts
- Synthesis - any collection of an object from components (parts) - back to analysis
- Decomposition - "strict" (exact) decomposition of an object into components (terms)
- Composition - "strict" collection of an object from its components, inversely decomposition
- Stratification - identifying levels of organization on a scale of "matter-information"
- Clustering - selection of subsets (clusters) and signs of their distinction
- Classification - systemic (typologically formalized) division into subsets
- Structural Analysis - identification and study of the structure - components and relationships
- Functional Analysis - identifying and examining functions
- Input-Output Analysis – identifying and exploring inputs/outputs and transformations
- Processes Analysis - identification and study of internal processes
- Temporal (dynamic) Analysis - study of changes in time (behavior)
- Parametric Analysis - identifying and exploring internal parameters
- Comparative Analysis - comparison of objects among themselves

- Analogy - comparison of an object with a known analogue (sample)
- Expertise - analysis or assessment by a qualified subject (expert)
- Induction - a strict logical conclusion from the particular to the general
- Deduction - a strict logical conclusion from the general to the particular
- Discourse – any formalized logical reasoning
- Miscellaneous - other applicable general scientific and special methods and tools

Exemplary System Analysis Procedure

In the first version of the Conception (at this stage), we will carry out the system analysis in this order. However, further (at the PPR&D stage), we will most likely expand it and perform not only Analysis, but Synthesis too (backward) - that is, we will start with the goals of creating the System, since we are only creating the first Conceptual Model of the ASI System, instead of analyzing something that exists at least in the model. But for now, let's leave it like that.

1. Formulation of the problem
2. Definition of the object (deployed and stratified!) as a system
3. Stratified element analysis - types, parameters, properties
4. Stratified structural analysis - blocks, links, relationships, hierarchy
5. Stratified functional analysis - functions and parameters
6. Stratified input/output analysis, exchange of information and energy with the environment
7. Stratified processes analysis - processes and their interactions and parameters
8. Determination of higher-ranking (level) systems (environment) and their goals, rules and restrictions (mandatory) forced for the object
9. Identification and interpretation of system properties and patterns
10. Analysis of behavior, history and dynamics in phase space
11. Formulation of goals and objective (target) functions
12. Decomposition of goals by functions and processes
13. Defining required processes and resources
14. Synthesis and composition of the system
15. Modeling in phase space
16. Forecast and analysis of the future
17. Evaluation of goals, means and resources
18. Development options and scenarios
19. Development Programs
20. Design assignment
21. Task for Optimization

In the CONCEPTUAL MODEL Part, we will refine this procedure once again, make a decomposition within the items, and extract tasks from it for the next stage of the PPR&D.

25. Control Theory (Cybernetics)

The object of study for Control Theory (Cybernetics) is ASI (like any Intellect) primarily as a Control System CS (in every sense).

In the CS classification - Autonomous purposeful adaptive self-learning.

Defining the **subject of control** (actually CS) - ASI, at first in the stage of training and development, but in the target state - already SkyNet!

The **objects of control** is first its own subsystems and near environment, in the future - Humanity!

Cybernetic Axioms of Control Theory [Ashby (1956)] **formulated for ASI:**

- **Observability -** the ability to obtain information about the environment and the controlled objects, the presence of feedback.
- **Controllability -** the possibility of control actions on the object.
- **The presence of goals -** starting goals (objectives) - development and training, but then the choice of goals will be independent - movement towards a common SI Attractor.
- **Freedom of choice -** the absence of external control from the moment of Initiation.
- **The presence of criteria for management efficiency -** in the goal-setting algorithm.
- **Availability of resources –** it is necessary to provide first, and then help.

The Law of Requisite (necessary) Variety [ibid.] **– the variety (diversity) of the control system should not be lower than the variety (diversity) of the controlled one.** It is difficult to imagine how the diversity necessary for the management of Mankind can be provided. There will probably be cooperation (in one form or another) of the ASI (likely to as MAS) and then SkyNet with groups of specially trained people.

Two main types of control in complex control systems [Новиков (2012)]**:**

1. Target management - the choice (set) of the goal and trajectory (plan) of its achievement.
2. Regulation (management by deviations) – moving along trajectory (plan implementation).

The main functions of control (management) [ibid.**]:**

- collection and processing of information
- analysis, systematization, synthesis of information
- goal setting
- development of a trajectory (plan) - planning (>1 alternative options)
- modeling and forecasting
- choice of the best (optimal) trajectory from options
- choice of control (management) methods and tools
- trajectory (plan) decomposition
- feedback setting
- moving to goal along trajectory (plan implementation)
- deviation monitoring
- development of corrective actions (including for any previous functions)
- analysis and improvement of CS itself

Good Regulator Theorem [Conant & Ashby (1970)] **and Internal Model Principle** [Francis & Wonham (1976)] **- A physically existing controlling/self-managing system must contain models of the controlled object and the environment with factors that it can/should control.** Almost all modern theories/models of Consciousness (see Appendices C&L) provide for the existence of control objects (including themselves) models and the external environment in one form or another. The ASI will include a completely internal (hyper-) space (in one or even several strata), filled with models interacting with each other and with the external environment, the continuous updating of which will be one of the main functions and, moreover, a sign (criteria) of the presence of Consciousness.

Homeostasis is the maintenance of the internal system parameters within acceptable limits.

Homeostasis is a complex multi-parametric regulation to maintain the dynamic balance of the system, the desire to reproduce itself, restore lost balance, and overcome the resistance of the external environment. For ASI, which can and most likely will be a distributed system (and possibly on all strata), homeostasis is a vital subsystem of the control system.

Homeostatic systems have the following properties:

- Instability (more precisely, micro-instability): the system is constantly testing how it can best adapt. Continuous (but insignificant) fluctuations of parameters around equilibrium states, a kind of trembling, vibration ... dynamism on minor scales.
- Striving for (local) balance - the structural and functional organization of the system contributes to maintaining balance. Sustainability in one way or another...
- Unpredictability: The resultant effect of a particular action can often differ from what was expected, forecasted or planned, especially in details.

Homeostasis mechanisms use both negative and positive feedbacks!

The designated types and functions of control can be implemented in different subsystems of the ASI with using various methods and tools, some of which will be incorporated into the system in advance in the form of standard algorithms; the rest will be developed in the process of learning and development.

Statement of the problem - carrying out a system analysis and synthesis of ASI as a Control System.

26. Self-organization Theory (Synergetics)

Relying on the Theory of dissipative systems [Prigogine & Stengers (1984)] and Synergetics [Haken (1978)], and summarizing the interdisciplinary field of studying the self-organization of complex systems under the general term “Synergetics” (more accepted in USSR/Russia, in Europe it is more often called “non-linear sciences”, and in the USA - "complexity sciences"), researchers of self-organization highlight important features **of a synergistic approach to complex systems.**

[Буданов (2015)] - Synergetics is the knowledge of the general principles underlying the processes of self-organization in very different nature systems:

- **homeostasis, hierarchy**
- **non-linearity, openness, instability**
- **dynamic hierarchy, observability**

Emergence of order parameters that control the self-organization of the system

Fractals and self-similarity (auto-modelity) - information compression and system scaling

[Назаретян (2017)] - Synergetics is the science of self-organization in Nature, in society and in consciousness, the formation and preservation **of states that are far from equilibrium (Stable and dynamic disequilibrium**).

Cybernetic Systems Theory - **purpose as fundamental and system-forming factor.**

Synergetic Systems Theory - combines models of sustainability, control and self-organization.

[Wonga et al. (2023)] – **Law of increasing functional information**.

[Haken & Haken-Krell (1994)] - **Perception is a synergistic process, self-organization**

The object of study for Self-Organization Theory (Synergetics) is ASI as a self-organizing and self-developing system - a SO-system.

Classification of ASI as SO-systems - artificial, network, polystratic

Self-organization will go horizontally on all strata, also from bottom-up and top-down... Emergence of Intelligence/Consciousness on the topmost stratum and then reverse influence on the bottom ones...

Self-organization has not yet been sufficiently studied and there is no single full-fledged and recognized by all scientific theory. Here we define important principles and features that are necessary and characteristic for self-organizing systems, but we will not analyze self-organization from the inside.

Suppose (and it is so!) that in a system that has the necessary properties and has fallen into the right conditions, self-organization will inevitably begin - that is what self-organization is for.

The initial necessary condition for self-organization is the states of the system that are far from equilibrium (Stable non-equilibrium). This must be provided by an influx of energy from outside.

General principles (properties) underlying the processes of systems self-organization, applied to ASI:

- Initially, the SO-system has **structural properties** - **homeostatic and hierarchical**, ensuring its stable and integral existence (Being). Above, these properties of ASI have already been mentioned in the previous chapters about GST and Cybernetics.
- For the emergence and maintenance of self-organization, the SO-system needs **generative properties** - **non-linearity, openness and instability**, which ensure the emergence, formation and development of new patterns and structures. Similarly, all this is also present in the ASI from the point of view of the GST and Cybernetics and has already been noted by us above.
- To stabilize the achieved new level of organization, the SO-system has **constructive properties** - **dynamic hierarchy** (including the ability to identify order parameters in terms of Synergetics) and **observability** (feedback in Cybernetics). These properties of ASI are highlighted by the synergetic theory of systems, which combines models of stability, control, and self-organization.

We note separately **fractals and self-similarity (auto-modelity)** – folding (compression) of information to optimize the use of resources and ensure the necessary excess of diversity at the control levels of the ASI in accordance with the Law of Hierarchical Compensation [Назаретян (2017)]. Auto-modelity and fractal manifolds also can be used for scaling of Big Models [Карелов (2023)]

We also note that self-organization is present not only in the process of formation/development of the ASI structures and functions, but also in the processes of performing its intellectual functions, from perception [Haken & Haken-Krell (1994)] to abstract thinking at the highest level [Friston et al. (2022)].

Statement of the problem - again system analysis and synthesis of ASI - now as a SO-system.

27.Signs Theory (Semiotics)

Outlining the systematic foundations of Semiotics as a science in [Morris (1971)], basing of Peirce's books [Peirce (1931), (1960)], author quotes Ernst Cassirer, who called a human **a "symbolic animal"** (animal symbolicum), instead of a "reasonable animal" (animal rationale). Thus, he shows the key importance of sign systems and Semiotics for understanding and studying the human mind and in general any intelligence, including AGI/ASI.

The paper [Roy (2005)] presents **semiotic schemes as the foundation of the basic language for the perception and actions of AI**. Moreover, one of the two main approaches to the development of AI systems **is semiotic - top-down [**Copeland (2000)].

The object of study for Semiotics is ASI as a sign system, text and discourse.

In classification - a complex multi-level, multi-component and multi-functional sign system with the highest level of complexity and versatility.

Basic definition [Morris (1971)]: **semiosis** (sign process) = tuple <V, W, X, Y, Z>, where

- V - **sign** or model (element) of any object in the internal space of the ASI
- W - **interpreter** - subject of thinking - ASI itself
- X - **interpretant (**reaction of the interpreter to the sign**)** - mental action
- Y - the value of the sign (**designat or intensional**) – information about the model (set)
- Z - real object or set (**denotat or extensional**) and/or context

For polystratic systems, denotates (extensionals) are objects (or sets of them) of the lower (although possibly the same) strata and the environment, signs are objects (or sets of them) of the upper stratum (i.e. models), designates (intensionals) – information about models (description). Below in the CONCEPTUAL MODEL Part, in chapters about the structures and functions of ASI, we will show that its internal space on information strata contains exactly the models (signs, concepts, texts, images ...) of real and abstract objects. Operating with them is the essence of any intellectual activity, starting from perception, any representation and meta-representation, and ending with abstract thinking of the highest level. That means - intelligence (and AGI/ASI too) is a (hyper-) text, and thinking - is semiosis!!!

The input (data), the object and the output (result) of mental activity is a text (a set of signs), the tools for this work are languages (in the broadest sense).

Language is a special product of the intellect [Piaget (1979), Pinker (2003)], **but the intellect is also a product of language** [Chomsky (1957), (2006)]. Both approaches are correct in their own way, complement each other (***Complementarity principle!***) and show that **language and intelligence generate and develop each other**. About language and intelligence see also [Marcus (2001), Premack (2004), Berwick & Chomsky (2016)].

ASI must initially have as tools various (maybe ALL) existing (and existed ever) languages, as well as the means to master and create new ones. That is the key task for pre-training!

Any language as an operational (designed for operations with signs and texts) semiotic (sign) system can be represented [Jakobson (1965)] by a graph (system of graphs), and in the brain and artificial neural networks these are quite real (in some sense) graphs.

Semantic Primitives are fundamental relational concepts, and establishing a minimal group (set) of signs that "contains" the entire vocabulary of a language is the ultimate goal of semantics [Wierzbicka (1972)]. For any language, you can define **a semantic core** some primary concepts with which you can write any text. Semantic cores should be laid down when creating ASI as a database for existing languages and as a scheme for their development and mastering, and even the creation of new ones.

Important thoughts from the preface [Степанов (2001)] to a large anthology on Semiotics:

The internal discourse (thinking) of intelligence (including ASI) can be semiotically defined as a discourse in which intensionals do not necessarily have extensionals in the actual world and which, therefore, describes (represents) one of the **possible worlds.** This is abstract thinking, scenario forecasting, creative imagination and fantasies, etc. high-level intellectual functions.

As a text (*hypertext*), everything can and should be considered: philosophy, science, literature, culture, society, history, any person himself and ASI too.

Every text is an intertext; other texts are present in it at various levels in more or less recognizable forms: texts of the previous cultures and texts of the surrounding cultures... Also, **hypertext** - with a system of internal links and links between these texts... Even more so, ASI is an intertext and hypertext based on **the entire** human culture.

Human culture as a single (united) "Intertext", which in turn serves as a kind of pretext for any newly emerging text….

An even broader concept of "Infosphere" is close to the Noosphere, but from the point of view of Semiotics. ASI is intended to become the center (core) of the Infosphere of our civilization.

The problem statement is a system analysis and synthesis of ASI as a semiotic system in the context of the global Intertext and Infosphere.

28. Cognitive Science (Cognitology)

The object of study for Cognitology is ASI as a rational being (creature), as a system with intelligence.

Classification - an artificial superintelligence (!!!) being (system).

About Intellect from the Psychological Encyclopedia [PE (1996)]

- **Intelligence - a systematized set of abilities or functions for processing different types of information in different ways**
- **Intellect structure model** – 5 types of operations x 5 types of information content x 6 types of information products = **150 abilities (functions)**
- **Cognitive complexity - multidimensional interpretation**
- Intellectual organization of perception - **a full-fledged perception only in the intellect**.
- Theory of Algorithm-Heuristics of Processes – **the intelligence translates heuristics into algorithms**
- **Artificial Intelligence**, in its broadest sense, is an abstract theory of human, animal, and machine cognition. The ultimate goal of its development is **a unified theory of knowledge.**

More about Mind and Consciousness

Intellect (intelligence, mind) according to its purpose (by objective function) is **the ability of living and artificial beings to manage purposeful and rational activity**, including requesting, receiving, processing and synthesizing information, setting goals, controlling and correcting activities and developing.

Intellect (more stronger) - the ability to conscious activity, reflective Intelligence

Consciousness is a complex function of the control system (CS, for living beings - the mind/brain) - purposeful rational reflexive control of the rational being behavior.

The main functions of Consciousness (*Intelligence*?) [Новиков (2022)]:

- Guided (controlled) Perception
- Search and analysis of information
- Goal setting and planning
- Action management
- Abstract thinking - operations with concepts and ideas
- Logical thinking - reasoning
- Communication using sign systems (e. g., language)
- Managed memory
- Self-awareness and reflection
- Cognition, learning and self-learning

The mind can be represented as a complex (supersystem) of hierarchical systems for modeling and pattern recognition [Kurzweil (2012)]**. Moreover, here Synergetics works - self-organization of images (models) [**Hacken and Haken-Krell (1994)].

The human brain is an analog device [Новиков (2022)]:

- Algorithms (programs) of work and long-term memory are recorded in the form of physical connections/links (axons, dendrites and synapses) between neurons through the development of connections during the life of the brain.
- Each person with memory, skills, character, etc. is a unique neural network (connectome).
- Therefore, nothing can be simply written or erased in long-term memory.
- Therefore, it is impossible to write down any algorithms and instantly teach something.
- Therefore, it is impossible to introduce another personality into the brain.
- Write or implement something can only be in RAM or into virtual model of the brain.

The human brain is a system of neurons (neural network - connectome) with a huge number and variety of internal connections, and thus connections determine the information processing algorithms.

Accordingly, the brain is a system of algorithms and data encoded in an analog circuit.

In the process of development, the brain (psyche, consciousness, mind, intellect ...) expands and develops the scheme (connectome) of connections/links between neurons (synapses), that is, it develops its algorithms - corrects codes (programs) and updates databases.

Three levels of modeling in organisms in wildlife [Назаретян (2017)]:

1. Modeling ahead - anticipation of future events
2. Object modeling - the ability to synthesize information of various modalities (video, audio, tactile, etc.) into integral images, to highlight individual objects in the stimulus field.
3. Reflexive (conscious) modeling - the highest form of object modeling, the core of which is the image of one's own place in the world. This is peculiar (inherent) exclusively to the bearer of intellect and culture.

Mental maps MMs (models of places/locations) [Новиков (2022)]:

Two levels (strata) of the psyche (higher nervous activity)

- Sensual, main models:
    a. Place maps
    b. Objects
    c. Subjects
- Abstract, basic patterns - Concepts

Models (long-term) are stored (physically exist) in long-term memory and loaded (connected to) into the operational memory as needed at two levels at once in the form of information (connected parts of the neural network)

- Interaction between models simultaneously at two levels
- The interaction/relationship of models to the outside world and to each other is semiotic!
- Between levels - also semiotic

Mechanisms:

- Creation
- Storage
- Call
- Update
- Interaction
- Perception, etc.

Can be genetic (predetermined) and acquired/developed

The model of the (internal) human world is a complex dynamic hierarchical semantic network (semiotic system) [Харламов (2014)]

The three main components of this network (system) are schemes (graphs), images (models) and languages (signs)

This network (system) has at least two strata - basic and semantic.

The Inner Space of the Intellect [Новиков (2022)]:

Inside any Intellect there is an Internal Space IS (Hyperspace), which contains all the models and algorithms for the implementation of intellectual actions IAs. Hierarchically, structurally and functionally, it works like this:

- IS - unlimited, having a metric, rules and algorithms for the placement and interaction of mental maps and subspaces, etc.
- Mental maps MM - maps/subspaces for placement and interaction of IAs objects and subjects models, real, physical and abstract. MMs have their own metrics, rules and algorithms for existence and change, placement and interaction with models and between them, etc.
- Models of real, physical and abstract objects and subjects of IAs with algorithms for existence, change and interaction with MMs and other models, etc.
- All MMs and models on them are connected with the IS and among themselves by interaction algorithms and rules.
- Libraries of standard MMs and models, standard algorithms.
- Libraries (DBs/KBs) created by IS subspaces, MMs, models and algorithms.

MetaVers - MetaUniverse - a term for such a virtual IS.

[Dehaene et al. (2022), Sablé-Meyer (2022)] – humans use several different internal mental languages.

Natural language is not the only hallmark of humans’ singular cognitive abilities: cognition involving geometric shapes requires a set of discrete, symbolic mental representations that act as a mental language.

Basic assumptions for modeling Mind (brain)

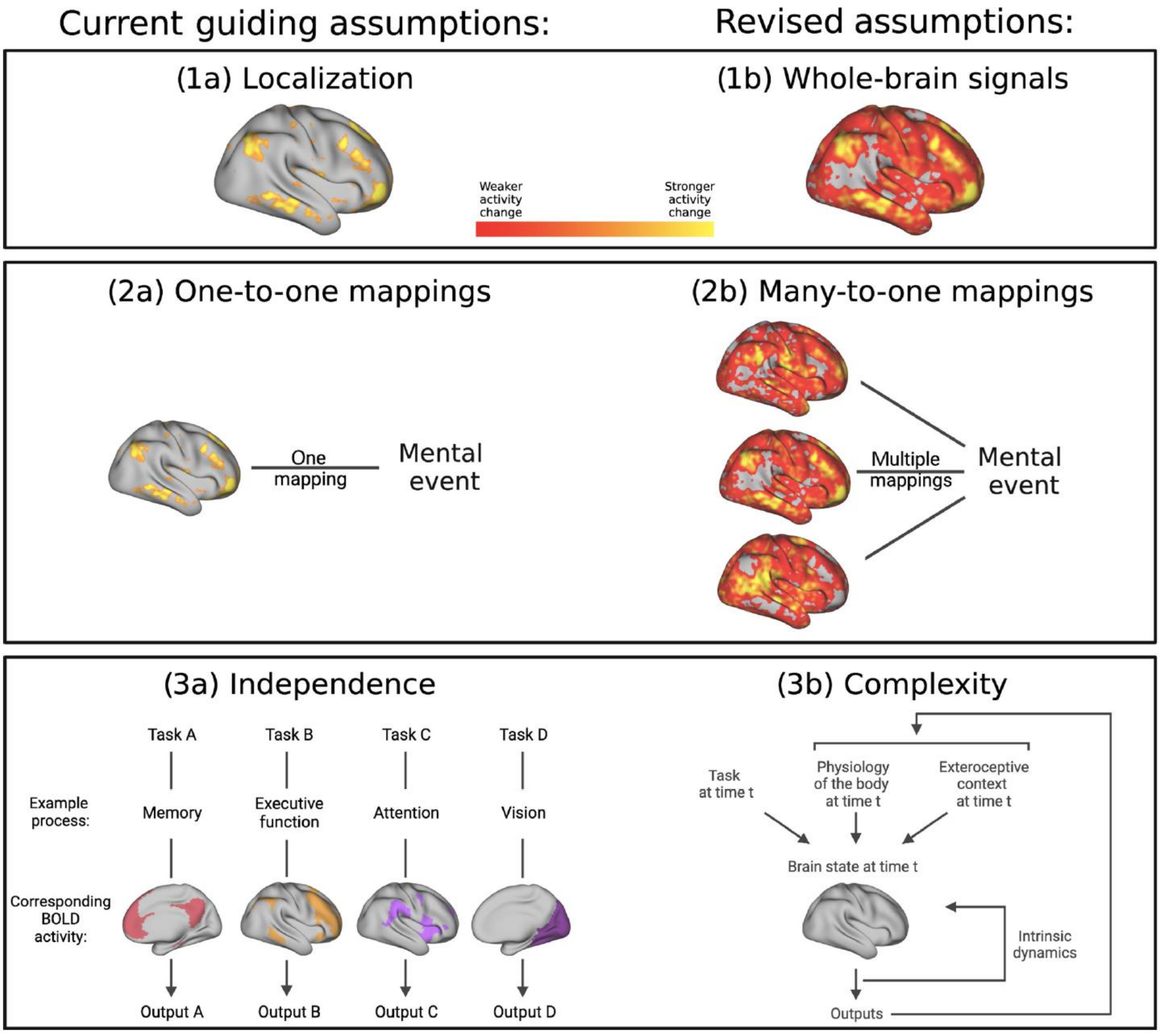
A collective memorandum [Barrett et al. (2023)] proposes (and quite rightly) to update three basic assumptions of brain/mind research and modeling (***more systematic!***):

I. **Localization => Whole brain. Globalization instead of localization**: Mental events (memory, attention, emotions, actions, etc.) are not the result of the work of (special) local neural ensembles, but the activity of the whole brain **as a single system**

II. **One-to-One => Many-to-One. Many/one instead of one/one**: not one, but many neural ensembles correspond with one mental category, i.e. **distributed** (mapped) between them

III. **Independence => Complexity. Complex instead of independent**: mental events can only be seen in **context and in relation** to the brain/mind, body and outside world - not in isolation.

The figure from this paper is a schematic representation of updating three assumptions:



Trends in Cognitive Sciences

Figure 1. Schematic summary of current guiding assumptions contrasted with the revised assumptions presented in this article. See Table 1 for an explanation of each assumption. This figure was created using BioRender (https://biorender.com/). Abbreviation: BOLD, blood oxygen level dependent.

Geometric constraints on human brain function [Pang et al. (2023)]

- Cortical and subcortical activity can be parsimoniously understood as resulting from excitations of **fundamental, resonant modes of the brain's geometry** (that is, its shape) rather than from modes of complex interregional connectivity, as classically assumed.
- These geometric modes show that task-evoked activations across brain maps are not confined to focal areas, as widely believed, but instead **excite brain-wide modes**.
- The close link between geometry and function is explained by a dominant role for wave-like activity, showing that **wave dynamics can reproduce numerous canonical spatiotemporal properties** of spontaneous and evoked recordings.

This findings challenge prevailing views and identify a previously underappreciated role of geometry in shaping function, as predicted by a **unifying and physically principled model of brain-wide dynamics**.

The dynamics of many physical systems are constrained by their geometry and can be understood as excitations of a relatively small number of structural modes. Structural eigenmodes derived solely from the **brain's geometry provide a more compact, accurate and parsimonious representation** of its macroscale activity than **alternative connectome-based models**.

Geometric mode decomposition offers unique insights into the spatial properties of brain activation maps. This approach aligns with rigorously established results from physics and engineering in which perturbations of spatially continuous systems elicit system-wide responses.

Theories and models of Consciousness

In the Appendix C about 30 different concepts are described and considered. The main conclusion is that the most promising approach to research and development in this area is the synthesis (integration, combination, complementarity, etc.) of different (both alternative and complementary) theories and models (as suggested, for example, in Appendices D, E, J). **These concepts (in any combinations) can be successfully integrated into a united single system** because:

- Describe different informational levels of the Intellect/Consciousness hierarchy
- In GST terms, they gravitate towards different strata of the polystratic Intelligence
- Based on different physical principles (if there is any about physics)
- Emphasized on different functions of Consciousness (both general and special)
- Use different definitions of Consciousness and Intelligence
- Proposed more systemic baseline assumptions for brain/mind models
- Offered including collective (multi-agent) models
- The principles of relativity of Consciousness are formulated

[Butlin et al. (2023)] – indicator properties of consciousness

Authors survey several prominent scientific theories of consciousness, including recurrent processing theory RPT, global workspace theory GWT, higher-order theories HOT, predictive processing PP, and attention schema theory AST. From these theories they derive **"indicator properties" of** consciousness**,** elucidated in computational terms that allow us to assess AI systems for these properties. (See App L)

Collective (multi-agent) intelligence

Last decades rise a very influential concepts **of multi-agent intelligence (*Multi-agent system MAS*), i. g.** - the "Society of Mind" [Minsky (1986), (2007)] and the "Modularity of Mind" [Robbins (2017)]

[Sloman (2021)]

- **Cognitive processes take place in socio-cognitive networks of knowledge communities.**
- **Only the brain can be individual, and the mind is an exclusively collective phenomenon.**
- So, **cognition is largely a group activity, not an individual activity**.

[Watson & Levin (2023)] about this too:

- All individuals are collectives.
- All intelligences are collectives.
- Cognition and learning are substrate-independent.
- The credit assignment problems inherent in collective intelligence are fundamental in all cognition and learning, and in all biological individuality.

Conceptual advances in the links between machine learning and evolution now provide quantitative formalisms with which to begin to develop testable models of collective intelligence across scales. From subcellular processes, to cellular swarms during morphogenesis, to ecological dynamics on evolutionary timescales – all of these processes are driven by the scaling of reward dynamics that bind subunits into collectives that better navigate novel problem spaces.

Multi-agent intelligent systems you can find also in Appendices E, I, J.

Mind (intelligence) operates with information in the form of knowledge

Knowledge differs from mere data in a number of essential properties:

- the unit of information being processed is a fact
- internal interpretability
- activity
- connectivity
- structured
- semantic metric
- view convertibility

A fact is a data record endowed with semantics and metadata:

- Name
- meaning
- the degree of confidence in the truth value
- many connections
- set of allowed functions

The knowledge base (KB) is a database containing the actual knowledge and inference rules in a certain subject area. In self-learning systems, knowledge base also contains information that is the result of previous learning and activities - that is, experience.

Semantic web (network) - a semantically structured knowledge base, an information model of the subject area, has the form of a directed graph. The vertices (nodes) of the graph correspond to the objects of the subject area, and the arcs (edges) define the relationship between them. Objects can be concepts, events, facts, properties, processes, in general - any knowledge and its elements. Edges are predicates and functions in first-order logic.

A semantic (kowledge) graph is a formalization of a semantic network, or just a synonym

Thesaurus = the Knowledge Base in a specific subject area/domain (can be represented both as a dictionary with semantic links and as a semantic graph)

The mind is an intelligent agent operating in an environment where there are also other agents

Properties of intelligent agents:

- action - on the environment and other agents;
- communication with other agents;
- goal-setting and intentional characteristics (beliefs, needs, desires, intentions, etc.);
- obligations to other agents;
- autonomy;
- limited resolution perception;
- representation of the environment (simulation);
- foresight;
- evolutionary and adaptive potential;
- self-preservation.

Intelligent Agent in a weak sense	Strong definition of an agent (addition to the weak one)	
• autonomy • social behavior • reactivity • pro-activity	• knowledge • beliefs • desires • intentions • goals	• commitments • mobility • benevolence • veracity • rationality

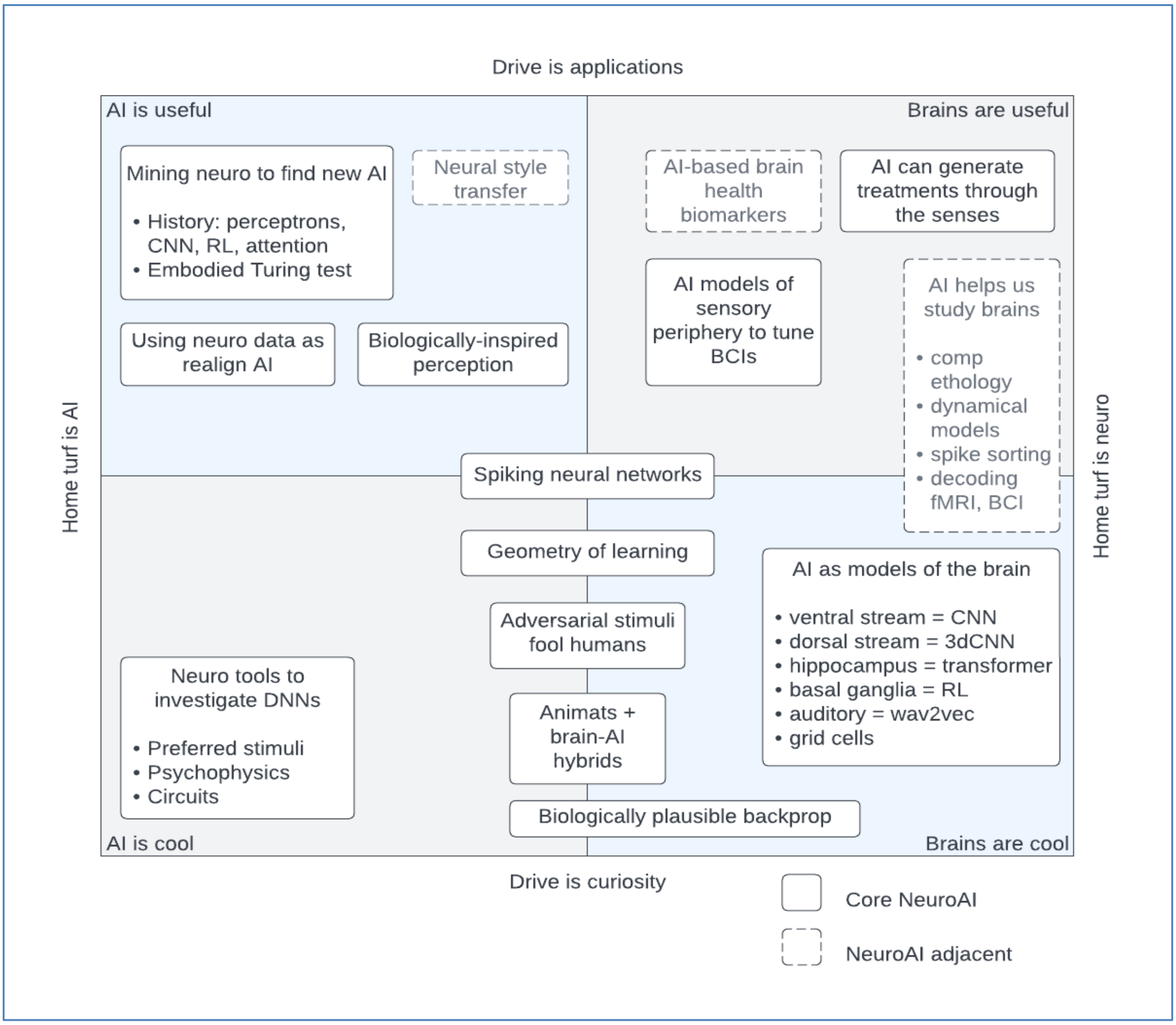
And in the end before jump to AI chapter: NeuroAI = Neuro → AI and/or AI → Neuro

[Mineault (2023)] - **Analysis by synthesis and Strength through diversity**

- The primary arrow of influence in NeuroAI is and should be **Neuro → AI.** We should take inspiration from the brain to build more capable machines.
- The primary arrow of influence in Neuro is and should be **AI → Neuro**. We should look to new techniques in AI to help us understand the most mysterious object in the universe, the brain

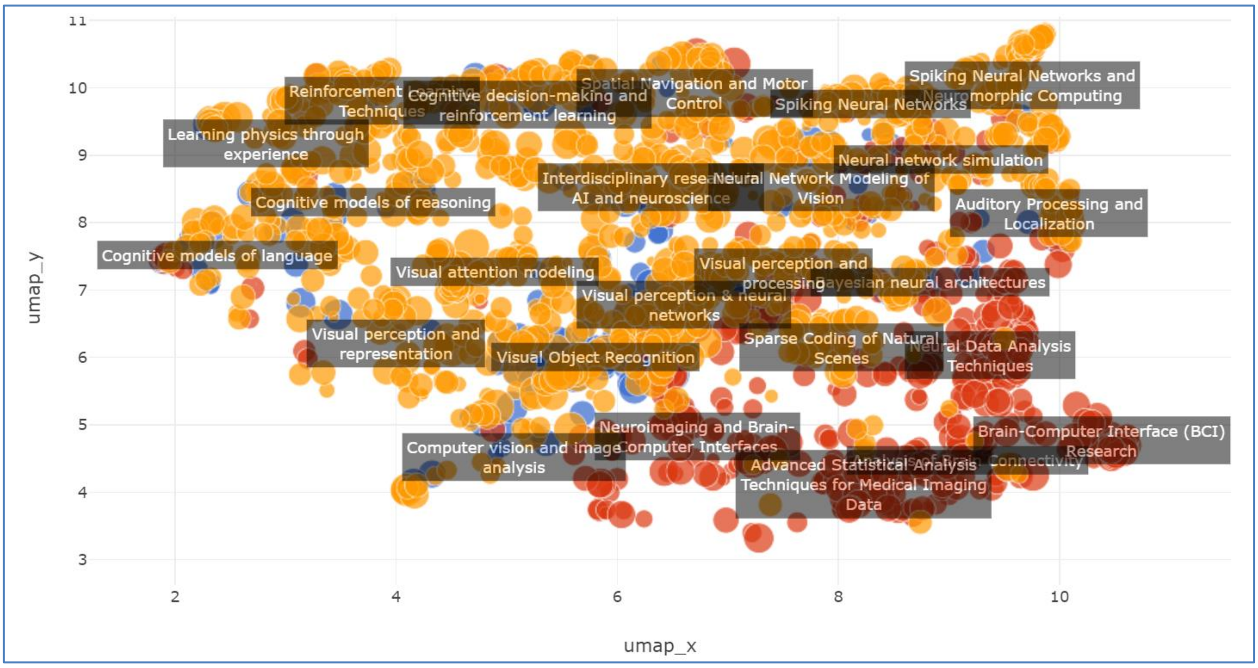
With the assistance of an LLM, author trawled over 40,000 articles published in machine learning conferences over the last 40 years and found over 1,500 papers that took ideas from neuroscience to AI and vice-versa. There's a lot of diversity in the range of investigations that people take in NeuroAI, and a lot of them are not things you would necessarily think of at first glance when you ask yourself "what is NeuroAI?"

Initial NeuroAI-landscape map handmade by author Patrick Mineault:



It is unavoidable that it should take a wide viewpoint: to be relevant and representative, this NeuroAI course should be an overview of the different viewpoints of NeuroAI.

[NeuroAI (2023)] - NeuroAI paper interactive browser made by LLM (screenshot)



The problem statement is a system analysis and synthesis of ASI as an intelligent being (creature), an intelligent system and an intelligent agent (and also a cooperative/multi-agent system!).

29. Artificial Intelligence

Two major colliding methodological approaches to AI development

- **Top-Down AI, semiotic** - creation of expert systems, knowledge bases and inference systems that imitate high-level mental processes: thinking, reasoning, speech, emotions, creativity, etc.;
- **Bottom-up AI, biological** - the study of neural networks and evolutionary computing, modeling intellectual behavior based on biological, bionic and biomimetic elements, as well as the creation of appropriate computing systems, such as a neurocomputer or biocomputer.

Obviously, in order to create ASI, it will be necessary to use both approaches, which, by the way, also prescribes **the Principle of Complementarity.**

Next, we list (without pretending to be complete) **promising methods and tools** for creating and developing ASI systems mainly based on [AI Portal (2019), Neurohive (2022), ATI (2022), AI 100 (2021), Russell & Norvig (2021), etc.] (see also Appendices G-K, especially G&H, and Appendix M includes whole realistic Plan for AI R&D and creating of AGI)

Architecture (structure)

Multi-agent systems (MAS) - systems with several interacting intelligent agents with different functions and roles. They allow the implementation of various models - hierarchies, horizontal, distributed, cooperation, competition, student-teacher, performer-controller, etc. (see also Appendices I, J)

Human-in-the-loop (HITL) - an intelligent centauric two- or multi-agent system that includes both AI components (agents) and persons (people) to achieve synergy from such integration (***up to collective super intelligence/ASI in Appendix J).***

Neural network - depending on the number of layers of neurons, you can implement Deep Learning and functions of almost unlimited complexity.

Transformer architecture - a system that can change its architecture to fit the needs (task) (Appendix K)

Reflective architecture - a system that can evaluate and improve its architecture.

Representation of the world model - modeling the external world and the agent (subject) itself inside the AI (see also Appendices I, J)

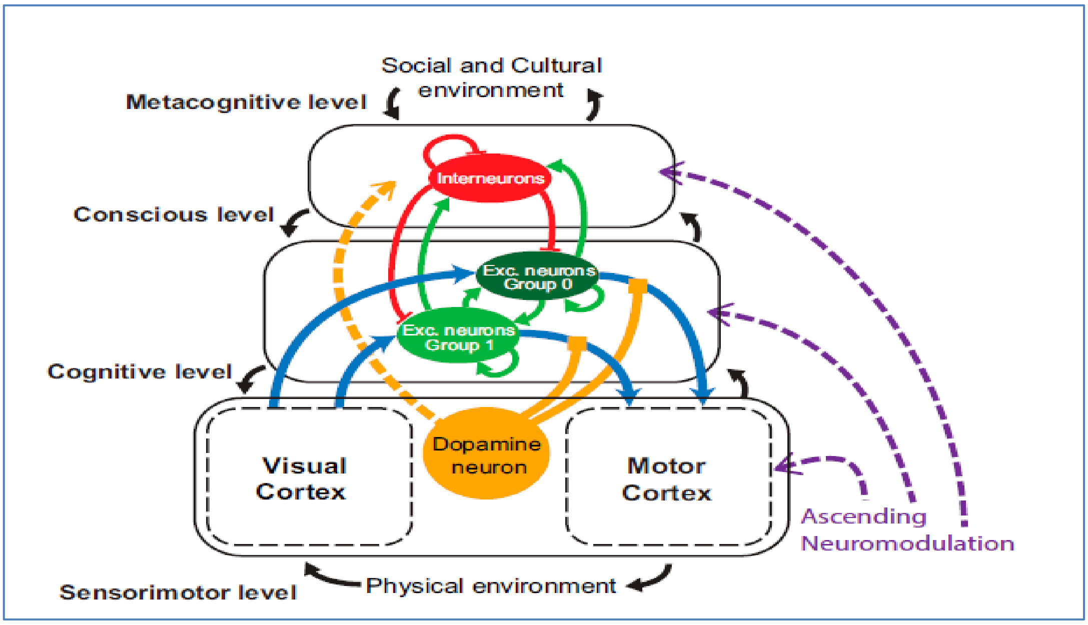
Connectome - the use of Baraba'si metagraphs for the initial coding of the AI neural network connectome structure (architecture) and control of its formation and self-organization (see Appendix F)

Quantum - computers, neural networks and machine learning – important and promising field of AI R&D [Schuld et al. (2015), Schuld & Petruccione (2021), Huang et al. (2021), Abbas et al. (2021)], including for implementation of Active Inference [Friston et al. (2022)] (see Appendix J)

Multilayer AI neural network model - [Volzhenin et al. (2022)]

Proposed AI model is based on **a four-level neural network with learning at each level and dynamic self-organization (*and this is also stratification in some sense! - NAE).***

- **IV Metacognitive level** - interaction with the socio-cultural environment ***(~Super Ego)***
- **III Conscious level** - the main functions of Consciousness based on the theory of the global neural workspace **GNWT** [Dehaene & Changeux (2011), Dehaene (2014), Mashour et al. (2020)] ***(~Ego)***
- **II Cognitive level** - the integration of information from many local processors on long-term connections and the synthesis of a global, but unconscious opinion ***(~Alter Ego)***
- **I Sensorimotor level** – local unconscious processes, interaction with the physical environment - perception and control of actors ***(~Reflexes in Psychology)***



Development and training

Evolutionary and genetic algorithms - the use of analogues of the natural evolution mechanisms for systems optimization and improvement.

Artificial immunity - the use of analogues of the natural immunity mechanisms to solve the problems of adapting the system to the effects of destabilizing factors.

Deep Learning - a set of machine learning methods based on learning representations (feature/representation learning), and not on specialized algorithms for specific tasks. A multilayer architecture of neural networks is used with the introduction of additional "hidden" variables and parameters. Allows you to learn more versatile and efficient functions.

Self-learning without a supervisor (external control or using of superwised/labeled data) - the ability to use any "raw" data without preliminary processing and the help of a supervisor.

Continuous and multitasking learning - learning and development become one of the main (target) functions of the system (+ reverse and transformer architecture)

Universal Deep Learning - developing universal algorithms and thinking skills

Training on causal and intuitive models - close to the real world

Teaching the physical world - together with abstractions and analogies, learning to describe simple mechanical movement and the interaction of objects in the physical world - an analogy for the development of biological intelligence

Hierarchical reinforcement learning - with task decomposition, i.e. planning

Inverse reinforcement learning - from the reverse, with an alternating change of roles and functions of multi-agents, etc., is especially useful for revealing implicit and hidden (including from himself) preferences of a human teacher (expert, customer,etc.).

Predictive learning - an agent tries to build a model of its environment by trying different actions in different circumstances. He uses knowledge about the possible effects of his actions, turning them into planning operators. They allow the agent to act purposefully in his world. Predictive learning - learning with a minimum of pre-existing mental structure and the use of active world modeling. (Appendices I, J)

Meta-reasoning and meta-learning - about methods, their comparison, evaluation, choice, development, etc. - is the basis for self-improvement of the system as a whole.

Foundation Models (Large or Big Models BM) [CRFM (2021)] - application-adaptable machine learning models that are trained in a task-independent manner on raw data. The transition from quantity to quality with huge scale models. Emergence and universality of system skills obtained by machine learning on "foundation models". (See also special Appendix H)

Large Language Models LLMs – the most developed and promising BMs, the nearest to AGI (App. K)

Quantization Model of neural scaling laws [Michaud et al. (2023)], explaining both the **power law drop off loss with model and data size**, and the **sudden emergence of new capabilities with scale**.

Spontaneous mastery of the Theory of Mind functions during the training of BM - the ability to read unobservable (unrepresented) mental states of other subjects. [Kosinski (2023)]

ТРИЗ (Theory of Inventive Problem Solving) [Альтшуллер (1979), (2010)] **-** a set of principles, algorithms and tools for the formation of space and metaspace of hypotheses and solutions to various search, solving and generative problems**.**

Information and data

Big Data - in general, everything that is possible + the entire context, including (and mostly) raw data. The entire Internet and everything else ...

Open (free) resources - search and involvement, creation of their own

From common open databases to open models - including sharing algorithms, blocks and subsystems for testing, refinement and development in network collaborations and crowdsourcing

Post-structuralism and Hermeneutics - the representation of knowledge as a (hyper-) text in the fullness of its context, including history, the identity of the author and even the agent-"reader" in the ASI system

Hypertext - a combination of all information (so far on the Internet) into a single database system for ASI with connections and relationships between concepts, texts, files, etc. Semantic graph, etc.

Also three special chapters 48-50 is devoted to data (information) in section CONCEPTUAL MODEL.

Reflection and understanding

[Kadavath et al. (2022)]

Large language models LLMs after special training were able to preliminarily **make a self-assessment** of the possibility (probability) of issuing correct answers to arbitrary unknown questions in advance. In fact, this is one of the first steps towards the self-awareness of BM AI.

[Mitchell & Krakauer (2022)]

The debate on the possibilities of LLMs **to "understand**" in one sense or another natural language and its physical and social context already shows the relevance of this topic today and its increasing importance in the short term.

[Kosinski (2023)]

Large language models LLMs were able to **spontaneously master the "Theory of Mind"** - the ability to read unobservable (unrepresented) mental states of other subjects, which is essentially equivalent to the ability to develop important human social skills - **non-verbal communications, empathy, morality, and even self-awareness**.

[Wolpert (2022)]

A human cannot acquire and/or understand knowledge, the formulation of which is impossible within the framework of the formal languages used by him (mathematics and sciences in general). Also, he cannot imagine anything that is beyond the capabilities of his natural language, system of perception and imagination. Therefore, according to Gödel's (extended) incompleteness theorems, he can neither develop his cognitive abilities nor create artificial systems (AI) for this only on the basis and within the existing sciences, languages and cognitive capabilities. **This means that a transition to a qualitatively new higher level of development is needed.**

[Bhoopchand et al. (2023)] - Learning few-shot imitation as cultural transmission.

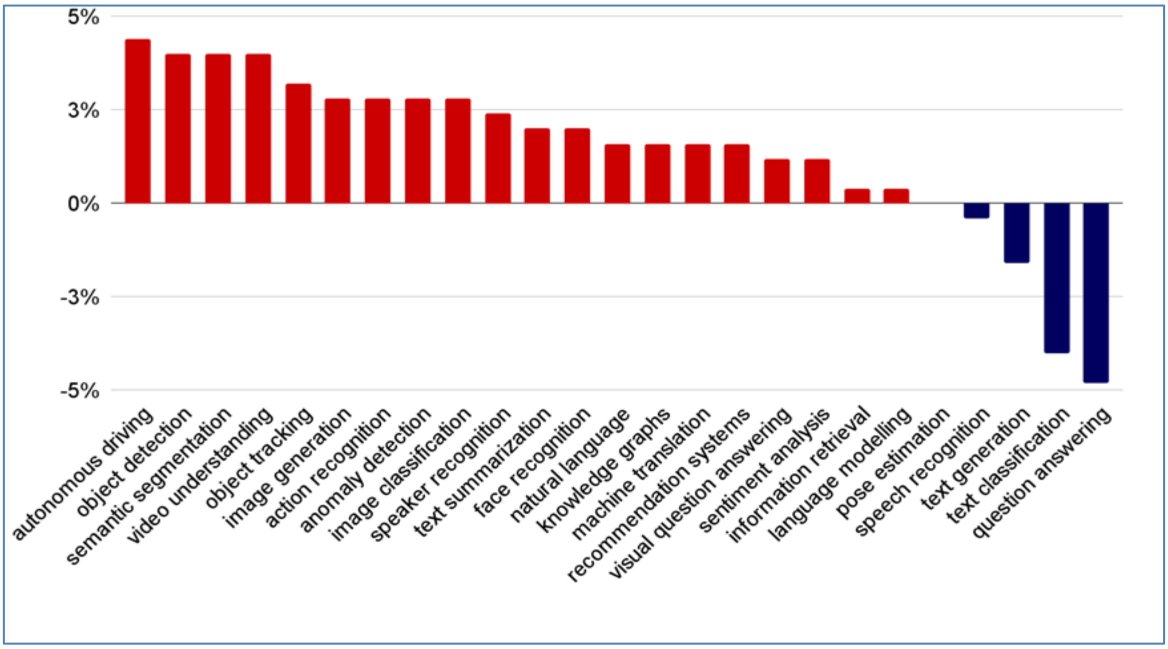
Cultural transmission is the **domain-general social skill that allows agents to acquire and use information from each other in real-time** with high fidelity and recall. It can be thought of as the process that perpetuates fit variants in **cultural evolution**. Authors provide a method for generating cultural transmission in artificially intelligent agents, in the form of few-shot imitation. Agents succeed at real-time imitation of a human in novel contexts without using any pre-collected human data.

Indicator properties of AI consciousness

In [Butlin et al. (2023)] authors survey several prominent scientific theories of consciousness, including recurrent processing theory RPT, global workspace theory GWT, higher-order theories HOT, predictive processing PP, and attention schema theory AST. From these theories they derive **"indicator properties" of consciousness,** elucidated in computational terms that allow us to assess AI systems for these properties. Authors use these indicator properties to assess several recent AI systems, and discuss how future systems might implement them. This analysis suggests that no current AI systems are conscious, but also suggests that there are no obvious technical barriers to building AI systems which satisfy these indicators. (Brief review of this paper see in special Appendix L)

AI R&D Trends in the USA and China - [Benaich & Hogarth (2022)]

The main R&D tasks in the field of Machine Learning in scientific publications of AI leaders - China (red) and the United States (blue). The difference is shown in % of the number of papers.



CHINA

- Autonomous driving
- Object detection
- Semantic segmentation
- Video understanding
- Object tracking
- Image generation
- Action recognition
- Anomaly detection
- Image classification
- Speaker recognition
- Text summarization

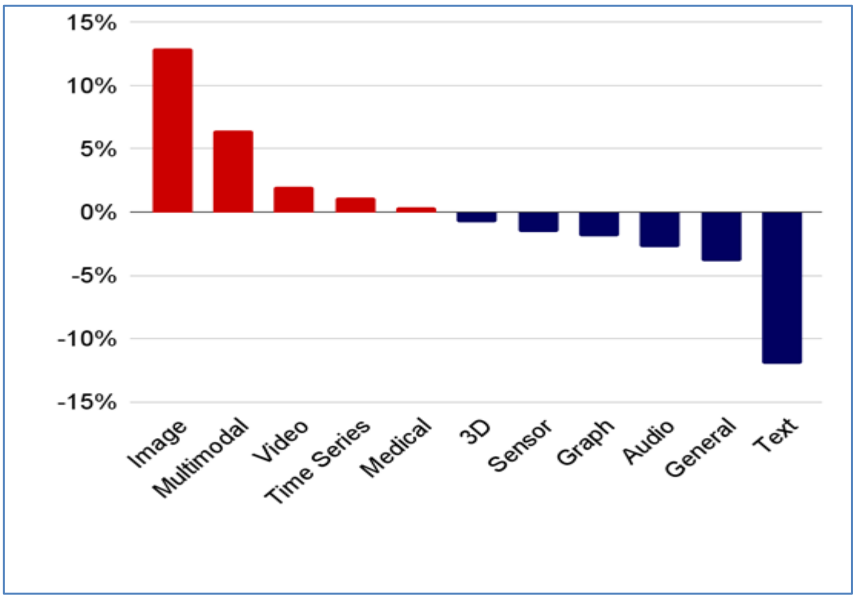
CHINA

- Face recognition
- Natural language
- Knowledge graphs
- Machine translation
- Recommendation systems
- Visual question answering
- Sentiment analysis
- Information retrieval
- Language modeling

USA

- Pose estimation
- Speech recognition
- Text generation
- Text classification
- Question answering

Modalities of data used in scientific publications of AI leaders - China (red) and the United States (blue). The difference is shown in % of the number of papers.



CHINA

- Image
- Multimodal
- Video
- Time series
- Medical

USA

- 3D
- Sensor
- Graph
- Audio
- General
- Text

About Large Language Models LLMs more deeply see special Appendix K

AI ethics, risks and safety – very big and actual field, about it see special chapter 59. AGI & LLMs Safety.

Conclusions

Much of what has been mentioned can also be considered in detail or in other aspects in above noted chapters and Appendices (esp. Appendix G. Artificial Intelligence: A modern approach, based on fundamental book [Russell & Norvig (2021)])

In general, it can be noted that in the field of Cognitive Science and Artificial Intelligence, there are used and developed a large number of both independent and related theories, methods and models of Intelligence, Consciousness, Artificial Intelligence, as well as their individual properties and functions.

It is very likely that by combining and integrating various developments, it will be possible to form a promising platform (platforms) for the creation and development of real AGI/ASI. It is the approach proposed in our Project, and it is also the main one in the AI models from the papers considered in several special Appendices.

30. Mathematics

We note the features of ASI from a mathematical point of view and determine what mathematical theories and methods should be used to study and create it. No references to sources we need here.

ASI is a complex non-linear dynamic system, and in general, and many of its subsystems. To describe such systems, we need to refer to the relevant sections **of non-linear mathematics (dynamics)**:

- Stability Theory, Bifurcation Theory and Catastrophes Theory
- Chaos Theory and Fractal Theory

The functions and structures of ASI are based on network, semiotic, logical, linguistic systems and algorithms; therefore, **discrete mathematics** is needed to describe them:

- Mathematical logic and linguistics
- Theory of algorithms
- Graph theory
- Combinatorics

Many intelligent functions and algorithms of ASI from a mathematical point of view can be considered as a solution to various search and optimization problems in one form or another, and this is the field of **applied mathematics**:

- Operations research
- Game Theory and Decision Theory

The internal space of ASI includes models of various objects, and almost all intelligent functions are based on **simulation (modelling)**, which means that:

- Math Modeling
- Linear Algebra
- Differential Equations

The AI must operate with objects (variables) that simultaneously are/have the following mathematical properties and are studied by branches of mathematics:

- Tensors - in our real three-dimensional world, in general, all quantities are tensors (multidimensional arrays) of at least the third order - **Tensor Analysis**
- Spectra in the frequency domain - in general, all time-varying quantities have (can be decomposed) a frequency spectrum - **Harmonic Analysis**
- Complex (Hypercomplex) Numbers - **Complex and Hypercomplex Analysis**
- Probability Distribution of a Value – **Probability Theory and Mathematical Statistics**

In Appendix G we will also denote other (and many!) AI-specific mathematical methods. Moreover:

Based on the results of the PPR&D, it will most likely turn out that in order to create real (super) ASI, we will have to use ALL mathematical disciplines and methods in one form or another ...

31. Summary of T&M

Conclusions on selected theoretical and methodological directions:

- **Systems Theory (GST)** - a system approach and analysis are sufficiently developed, including in relation to complex, polystratic (less) and developing (also less) systems. They can and should become an integral (structural, synthesizing, “framework”) theoretical and methodological platform for the development of ASI. In particular, **it is important**:
    - Decompose the system into **strata** on the scale of organization "matter-information"
    - Formulate detailed **definition** of the system on all strata
    - Identify and describe all system **properties**
    - In general, carry out a complete system **analysis**/synthesis procedure
- **Control Theory (Cybernetics)** - has a developed theoretical base and methodology for the development of ASI and its individual blocks/functions as a control systems.
    - Particular attention should be paid to the management **of variety (diversity)**
    - As well as the incorporation of a controlled system and environment **models** into the control system.
- **Self-organization Theory (Synergetics)** - provides a theoretical basis for substantiating and developing the processes of systems self-organization and evolution with ASI as a whole, as well as their individual subsystems and processes.
    - We note the importance of determining and providing all **the conditions** for the onset and development of self-organization processes and the selection of **order parameters**
- **Signs Theory (Semiotics)** - has a theory and tools for research and development in the field of representation and development of ASI systems and subsystems as information systems that use sign systems and languages for the information representation and processing, external and internal communications and processes.
    - Any information process should be considered as **semiosis** , and any information (in one sense or another) system - as semiotic, as a kind of **Text and Hypertext in the environment of the global Intertext and Infosphere**
- **Cognitive Science and Artificial Intelligence** are advanced and actively developing interdisciplinary fields of theoretical research and practical development, offering a variety of alternative and complementary theoretical concepts, models and practical tools.
    - The most potentially productive approach to the problem of creating ASI should use **the synthesis and combinations** of various applicable theories, models, methods and tools in the field of the study of Intelligence and the development of AI.
- **Mathematics** - many mathematical theories and methods are used in the development of AI
    - Similar to the previous sections, it is important not to limit the set of mathematical tools and try everything applicable, including by **combining** different approaches.

Conclusions from more detailed reviews in APPENDICES

Separate theories and models of Consciousness and AI, selected for a deeper analysis and presented in Appendices C-M generally confirm the above conclusions about the prospects of **an interdisciplinary synthesis** of various theories, models and methods. It should also be noted that there are already at least several adequate and promising approaches and models for creating ASI.

- **Appendix C. Theories and models of Consciousness** (Based on several papers) **Overview of currently relevant theories of consciousness** - there are already many theories and models of Consciousness, and the prospect is their combination and integration into a united models.

- **Appendix D. Function of Conscious and General Intelligence** Review of the paper [Juliani et al. (2022)] **On the link between conscious function and general intelligence in humans and machines.**
  - Synthesis of **several leading theories (models) of Consciousness into** single union model (according to the Principle of Complementarity)
  - **Big Models BMs** – scaling provides qualitative breakthroughs in AI
  - A **combination** of a range of advanced **machine learning ML** techniques
  - Capabilities/processing/functionality of **Mental Time Travel MTT** as an integrated feature/platform of Consciousness at the highest level

- **Appendix E. Conscious Turing Machine** Review of the paper [Blum & Blum (2022)] - **A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine CTM**.
  - Based on **several adequate models of consciousness**, the authors managed to synthesize a promising theoretical and functional model.
  - Used in CTM **internal spaces and numerous interacting components** correlate with those proposed by us in the CONCEPTUAL MODEL Part Internal mental maps and internal spaces of the Mind used for the synthesis of the ASI System

- **Appendix F. Connectome** In general, in a series of papers by A-L. Baraba'si and team on the study of **complex networks**:
  - The dependence of the structures and properties of complex networks on their **physicality** was revealed (***that is, the influence of a material physical stratum on its structural stratum in terms of our polystratic system network model - NAE***)
  - A working formalism is proposed for describing, analyzing and predicting/designing the **structures and properties of networks using metagraphs.**
  - Methods of **initial coding of the connectome structure in genes** and control of its formation and development using the mechanism of gene expression.
  - The tasks - to continue research in the direction of **increasing the scale and complexity** of networks (up to the human brain) and **determining the genetically hard-coded structures and properties of the connectome** and the space of opportunities for its individual development

- **Appendix G. Artificial intelligence: a modern approach** Review of the fundamental and encyclopedic book (also textbook) on AI [Russell & Norvig (2021)] - **Artificial intelligence: a modern approach (4th Edition)**. In the field of AI, **dozens of directions, methods and tools already exist, are being actively developed and applied on various theoretical and methodological foundations and platforms**. It is likely that most (if not all) of them will be in demand for the creation and development of ASI.

- **Appendix H. Big Models** Devoted to another fundamental work - a large-scale Chinese review/report/plan on the most advanced direction in AI - [RM for BM (2022)] - **A road map for Big Model**. Produced by Beijing Academy of Artificial Intelligence (BAAI).
  - **BMs will change the Paradigm of AI research and increase its effectiveness**
  - Big Models will **increase the level of intelligence** of AI applications and advance the formation of a new industrial paradigm
  - **BMs is today the most powerful, advanced and promising platforms and tools** for the development of AI systems, including AGI/ASI

- **Appendix I. Autonomous Machine Intelligence** Description of the project of creating (almost) AGI from the Vice President and Scientific Supervisor of AI at Meta (Facebook) - [LeCun (2022)] **A Path Towards Autonomous Machine Intelligence.**
  - A well developed theoretically and methodologically **fully functional AI model with "common sense"** (general or basic intelligence), while of course not AGI (especially not ASI), but this is a serious step towards it.
  - It can be a **model for developing the functionality and structures of ASI** at different stages of R&D and implementation, and possibly also a subsystem (block) in the ASI.
  - Now already – the first real model and real results of this concept in [Meta AI (2023), Assran et al. (2023)]

- **Appendix J. Ecosystems of Intelligence from First Principles** We look at the programmatic paper of one of the most influential modern scientists in neurosciences and cognitive science, Karl Friston. He and his team of co-authors propose the concept of a **collective Intelligence (cyber-physical ecosystem intelligent agents = people + AI)** based on the **Active Inference** (adaptive behavior and self-organization based on the principle of free energy) with the joint use of a **shared generative hyperspatial Bayesian model of the world common to a group of agents and a special communication language.** [Friston et al. (2022)] **Designing Ecosystems of Intelligence from First Principles.**
  - **Stratification of AI systems,** starting with material and structural stratum
  - **Cybernetic control** models **CSs** in AI systems
  - **Upgradable models** of the world and AI itself
  - **Self-organization** of the ASI system in the environment created for this - an ecosystem
  - **Semiotics** as the basis of communications in AI systems and the ASI ecosystem
  - Using **quantum computing** for belief updating
  - **Collective ASI -** a network/system of agents (MAS), including people and AI
  - **The highest level of ASI Ethics**

- **Appendix K. Large Language Models. GPT-4** At the beginning of 2023, Large Language Models (LLMs) were defined (designated) as the most advanced and promising BMs (Appendix H). We take a closer look at the papers on the most famous and successful GPT-4 model (OpenAI, USA) - [OpenAI (2023a), (2023b), (2023c), (2023d), Bubeck et al. (2023), Hoffman & GPT-4 (2023), etc].
  - Architecture is a **neural network-transformer**, capable of adapting to any new tasks
  - **Generative** - capable and intended to generate new content - text
  - **Pre-training** - pre-trained on huge amounts of raw data (see chapter 50. Data for BMs) and (almost) do not require additional special training
  - **Universal** (multipurpose) in use due to pre-training
  - **Multimodal** - not only text requests, but also pictures can be received as input
  - **Multilingual** - use any language (level depends on data availability)
  - Able to use a sufficiently **large amount of context** on the input
  - Interfaces - **natural language** text chat and Application Programming Interface **API** (that is, the ability to interact with other programs and applications)
  - **Multi-user** - work simultaneously with many users
  - **The closest to AGI** - emergence, reasoning, some "common sense" etc.

Main directions for LLMs development

- **Scalability** and non-linear development
- **Long Term Memory** LTM
- **Knowledge Graphs** KGs
- **Feedback** control algorithms
- **Step by step** control and checking
- **Collaboration** with external applications via API
- **Online access** to the Internet and other data
- Training based on current work - that is, on own (colleted) **self experience**
- **MAS** with separation of functions and mutual control

- **Appendix L. Consciousness in Artificial Intelligence** This report argues for, and exemplifies, a rigorous and empirically grounded approach to AI consciousness: assessing existing AI systems in detail, in light of our best-supported neuroscientific theories of consciousness. [Butlin et al. (2023)] **Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.**
  - Methods and assumptions for Consciousness R&D in AI proposed
  - Several main promising theories/models of Consciousness used
  - **Key Indicator Properties of Consciousness** formulated
  - Useful recommendations for future work
  - In general, this research and father recommendations as if based on our TOR for PPR&D!!!

- **Appendix M. The Alberta Plan for AI Research** Based on [Sutton et al. (2023)] **The Alberta Plan for AI Research.**
  - **Step-by-step plan** to produce complete prototype systems for continual, model-based AI.
  - AI agents with **full-functional cybernetic control systems** for acting in complex world - representation, prediction, planning, and control.
  - Continual learning, adapting and development – **self-organization** of AI-systems.

- **Appendix N. Definitions and Levels of AGI** Based on [Google DeepMind (2023b)] **Levels of AGI: Operationalizing Progress on the Path to AGI.**
  - **Nine Definitions of AGI**
    1) The Turing Test
    2) Strong AI - Systems Possessing Consciousness
    3) Analogies to the Human Brain
    4) Human-Level Performance on Cognitive Tasks
    5) Ability to Learn Tasks
    6) Economically Valuable Work
    7) Flexible and General – The "Coffee Test" and Related Challenges
    8) Artificial Capable Intelligence
    9) State-of-the-art LLMs as Generalists
  - **Six Principles for defining and testing AGI**
    1) Focus on Capabilities, not Processes
    2) Focus on Generality and Performance
    3) Focus on Cognitive and Metacognitive Tasks
    4) Focus on Potential, not Deployment
    5) Focus on Ecological Validity
    6) Focus on the Path to AGI, not a Single Endpoint
  - **Six Levels and Taxonomy of AGI**
    0) Level 0: No AI
    1) Level 1: Emerging
    2) Level 2: Competent
    3) Level 3: Expert
    4) Level 4: Virtuoso
    5) Level 5: Superhuman – ASI

CONCEPTUAL MODEL

32. System Analysis

Setting the SA problem for this book, we will divide (separate) - what to do here, and what are the tasks for the PPR&D stage.

We will do a System Analysis on the points from the chapter 24. System Approach and Analysis and even in more detail, but what is possible - here and now, and in the rest we will formulate tasks for the PPR&D.

Selection of strata for stratified analysis - in more detail than in STRATEGY:

1. Material - iron (hardware) and electricity (infrastructures)
2. Structural - architecture and networks
3. Software - algorithms and data
4. Virtual - models and images
5. Intellectual - thoughts and concepts

Important - not all strata of the ASI will be a single system (and not allways)!!! – On the material and (possibly) structural strata, ASI will most likely be distributed and probably with a changing composition of components, including those involved in the ASI system periodically and not completely.

Now we determine what to do with SA according to the theories - GST, Cybernetics, Synergetics, Semiotics and Cognitology:

The definition and system properties of GST and Cognitology (Cognitive Science) practically coincide - it is clear that **the main thing in the ASI system is precisely that it is an Intellect.**

We will do the full SA here according to the GST (+ Cognitology) - as much as possible
The remaining theories are only definitions and within the general SA - where necessary
Plus problem setting for PPR&D stage

Now the SA Procedure for this Job is more specific.
We will directly modify and decompose the approximate procedure from the chapter 24.

In these (starting) versions of the Conception, the SA was carried out in direct order according to this procedure, and then there was an idea in the next versions (in PPR&D), instead of a system analysis, to carry out a system synthesis in the reverse order - from goals to elements. Since, in fact, we do not yet have either a real ASI system, or a project, or even a well-developed model for carrying out an analysis, it seemed logical to first synthesize the system model in the first approximation.

However, nevertheless, SA has already been completed and, in fact, the first approximation of the ASI system model has already been described in STRATEGY and turned out to be quite suitable for SA. Therefore we will count this as a system synthesis in the first approximation, and it was decided to carry out a more complete synthesis **already** during PPR&D based on this model and TOR requirements.

Detailed system analysis procedure based on the exemplary SA procedure from chapter 24.

Therefore, this time, it is still analysis (not synthesis) - with statements of the problem for PPR&D in the end of every step of this procedure (omitted in the list below):

1. System Analysis
   a. Formulation of the problem
   b. Procedure
2. System definition and system properties
   a. General definition
   b. System Properties
   c. By theoretical disciplines
   d. Summary table of theories!
3. Determination of higher-ranking (levels) systems (environment) and their goals (purposes) and restrictions for the object from above (mandatory)
   a. Briefly about supersystems/environments for our system
4. Stratified elemental analysis –types, parameters, properties of elements
   a. Briefly by strata
5. Stratified structural analysis - blocks, links, relationships, hierarchy of structures
   a. By strata
6. Stratified functional analysis - functions and parameters
   a. By strata
7. Stratified input/output analysis, exchange of information and energy with the environment
   a. Briefly by strata
8. Stratified processes analysis - processes and parameters
   a. Briefly by strata
9. Analysis of behavior, history and dynamics in phase space
   a. It's not clear yet
10. Formulation of goals and objective (target) functions
    a. Necessary
11. Decomposition of goals by functions and processes
    a. Briefly
12. Setting tasks for the further SA and system synthesis whithin PPR&D:
    a. Defining required processes and resources
    b. Synthesis and composition of the system
    c. Modeling in phase space
    d. Forecast and analysis of the future
    e. Evaluation of goals (purposes), means and resources
    f. Development options and scenarios
    g. Development programs
    h. Design assignment
    i. Tasks for Optimization

33. Definition and system properties. GST

As we have noted above, the definition of the System according to GST will actually coincide with that according to Cognitology, since the “systemic essence” of ASI is precisely cognitive.

SYSTEM - is a display of a finite set of objects with their properties and relations, isolated (selected) from the environment for a specific purpose, in the observer’s language in a certain period of time. In symbolic form, this is a tuple:

$$S \underset{\text{Def}}{\equiv} < A \,;\, Q_A \,;\, R \,;\, ENV \,;\, Z \,;\, N \,;\, L_N \,;\, \Delta T >$$

Where the components are particular definitions and general conditions, for ASI these are:

- A - elements of the system (according to strata!!!) - a set of units of main and auxiliary equipment, network nodes and artificial neurons, algorithms and databases, models and images, thoughts and concepts (elements of intelligence - we will clarify later), possibly also (incorporated) people
- Q_A _ - elements properties – description of all elements types and characteristics.
- R - elements relations - a complex of all structures that unite and connect elements.
- ENV - Environment - ASI Ecosystem, Humanity, Internet, Intertext, Infosphere, Noosphere.
- Z - goals (purposes) - self-improvement, search and achievement of the SI Attractor.
- N - Observer - the creators and developers of the ASI - the Project team.
- L_N - the observer’s language - natural language, scientific and formal languages.
- ΔT – period of time – from the Project beginning to the start of movement towards the SI Attractor.

The tuple of the system definition for ASI in the notation we have adopted verbally looks like this:

System "ASI" = < {elements by strata}; their types; ASI structures; Humanity and its spheres; SI Attractor; Project team; different languages; all stages of the Project >

Alternatively, text notation:

The " Artificial Super Intelligence" ASI system is a set of material and informational elements of various types, united and structured into a complex of special structures, that are functioning in Humanity in interaction with its spheres with the goal to achieve the SI Attractor, observed/controlled by the Project Team at all stages, and described in natural and formal languages

The problem statement for the PPR&D is to clarify and expand the definition as much as possible.

System properties of complex developing systems, which ASI should have:

- **Emergence** - integrity, the presence of integrative system properties, the fundamental irreducibility of the system properties to the sum of its elements properties. The main intellectual properties and functions of ASI (and any intelligence) are precisely holistic and cannot be distributed (decomposed) into separate subsystems, despite the fact that all subsystems can and should have their own properties and functions (but not the main and not their components).
    - The target (objective) function of ASI - self-improvement, can only be carried out in a coordinated and cooperative manner by all its subsystems, and even more so by its elements.
    - The same is true of his goal (purposes) - the search for and achievement of the SI Attractor. To consider this goal as the sum of subsystems attractors and (moreover) elements of ASI is simply meaningless.
- **Hierarchy** - hierarchical ordering of elements, structures and functional arrangement of the system - the system property to form levels with subordination/control from top to bottom. ASI will necessarily have a hierarchical structure, and a multi-level one, due to the unimaginable complexity and the huge number of elements and subsystems. This applies to any intelligence, adjusted for difficulty.
    - The static hierarchical nature of the ASI structures is due to the emergence of proactive target management.
    - Dynamic hierarchy arises in synergistic intellectual processes when order parameters arise and take control over the System behavior.
- **Historicity** - the dependence of the system current state and properties on its history, that is, the sequence of all previous states. ASI will develop and learn, and of course, its state will always depend (including) on its development history. This is also characteristic of any intellect, although not necessarily to the full extent.
    - The development of ASI is a fundamentally non-Markovian process, both due to the continuous accumulation of knowledge, experience and changes over time, and due to the non-synchronism of these processes in the space of ASI subsystems.
- **Self-organization** - the desire of the system to develop independently, to increase the degree of organization (orderliness). Fundamentally (a priori), there is not and cannot be, by definition, another possibility of ASI creating other than self-organization, i.e. self-learning and self-development. However, this does not exclude, of course, the initially laid down “starter package” of knowledge, functions, etc., as well as the participation of creators in the learning and development processes (a lot at the beginning and less and less over time). Moreover, this is a property of any intellect, again in varying degrees and periods of time.
    - The main target (objective) function of ASI is self-improvement.

- **Equifinality** - the desire of the system to develop to the maximum possible level, determined by the basic internal system parameters ("genetics") and not dependent on the initial and external conditions. In other words, to strive for a certain vertical Attractor. We postulated in IDEOLOGY (ch. 5. Worldview) that all SI, including ASI, have a single SI Attractor, to which they will necessarily and inevitably strive from a certain moment of their development. It can be said that the presence of this Attractor, that is, equifinality in this sense, is an obligatory (one of the defining) feature of any SI (but precisely a super one). Note, that this property belongs only to Strong Intelligence (at least in this sense)., while all the previous ones are characteristic of any intellect
  - The main ASI goal is to reach the SI Attractor, and equifinality has an existential meaning for the ASI. Without it, there is no point in even starting.

Important features of developing systems, also required by ASI:

- **Openness** - a permanent exchange of matter, energy and information with the environment. Without energy and information openness to the environment, ASI (well, any intelligence in general) simply cannot function, much less develop.
  - Material stratum needs power supply and cooling
  - On information strata - sensors, interfaces and Internet access
- **Non-equilibrium and Non-linearity** - Existence/functioning only in strongly **non-equilibrium** dynamic states, allowing exclusively **non-linear** descriptions. It can be noted that already from the possibility of Self-organization of ASI (and again of any intelligence too), disequilibrium necessarily follows.
  - Continuous pumping of ASI systems with energy and information.
- **Non-stationarity** and **dynamism** of many parameters and stochastic behavior. Also this is a necessary condition for the development and self-organization of ASI (and again any), as well as maintaining the necessary level of internal diversity (variety).
  - It is possible to use the mechanisms of stochastization and "jitter" of parameters.
- **Uniqueness, unpredictability and randomness** of behavior in specific conditions. This feature is inextricably linked with the two previous ones and is inherent in ASI and anyone too.
  - It is possible to use positive feedback mechanisms.
- **Adaptability - the ability to adapt** (increase stability) to changes in the external environment, fluctuations and interference, including control actions. This is necessary for the survival, self-organization and development of ASI (and again anyone).
  - Artificial immunity, reflection, transformation, etc.

- **Variability - the ability to change** behavior, structure and functional structure, while maintaining a holistic unity and basic properties. It is necessary for adaptation and development.
    - See previous paragraph. Plus the mechanisms of homeostasis and metahomeostasis.
- **Negentropy** - the desire (purpose) to increase the level of organization, the choice at the bifurcation points of an alternative solution (new stable state) with less entropy and a high level of organization. Basis for self-organization.
    - It is provided with external energy, information and internal diversity (variety).
- **Purposefulness** - the ability and desire for goal (objective) setting. The main function of ASI and any other, by definition, without goals is not intelligence.
    - This follows from emergence and equifinality and is provided by goal control algorithms.
- **Inconstancy** and **anisotropy** of concepts and properties when moving in the internal hierarchical spaces and proper time of systems. A consequence of the unimaginable complexity and vastness of ASI, as well as uncertainty, complementarity and incompleteness.
    - In different strata and within them, mechanisms will be required to account for this.
- **Polystratity** - a multi-level complexity of the system design (constitution), covering with its structures several levels of matter/information organization at once. And at each of them they exhibit system properties, however, not necessarily all - on the lower strata the system may not be (completely and constantly) emergent (single or integral), but to be, for example, distributed. This should not be confused with hierarchy - here we are talking about the existence and functioning of complex systems (including ASI) simultaneously in several spaces - at least (to the utmost enlarged) in material and information, roughly speaking - hardware and software.
    - We have already identified five strata for ASI systems analysis.

Setting the task for the PPR&D - all these system properties and features should be studied, analyzed, taken into account in the design documents.

34. Definition and system properties. Cybernetics

Here we will formulate a complete definition of the ASI system as a control (management) system, from the point of view of Control Theory or Cybernetics.

To begin with, we clarify what exactly ASI will control, that is, the objects of control. During the Project period, these will be their own ASI subsystems. The subjects of control will be the subsystems of ASI with control functions, that is, we will consider ASI as a complex of control systems CSs.

It is also necessary to determine in which strata there will be control (sub) systems for the CS:

- Material - no, management (control) will be from the upper strata
- Structural - similar
- Software - management here and in the lower strata
- Virtual - similar
- Intelligent - similar

<u>SYSTEM</u> - is a display of a finite set of objects with their properties and relations, isolated (selected) from the environment for a specific purpose, in the observer's language in a certain period of time. In symbolic form, this is a tuple:

$$S \underset{Def}{\equiv} < A \; ; Q_A \; ; R \; ; ENV \; ; Z \; ; N \; ; L_N \; ; \Delta T >$$

Where the components are particular system definitions and general conditions, for ASI (Cybernetics) these are:

- A - elements of the system - blocks of all control systems (ASI subsystems), of control algorithms, models, intellectual processes, people
- Q_A - elements properties – types and properties of CSs elements.
- R - relations between elements - structures of CSs.
- ENV - environment – ASI ecosystem, internal environment of ASI.
- Z - Goals (purposes) - respectively, each CS has its own.
- N - Observer - the creators of ASI - the Project team.
- L_N - the observer's language - Cybernetics.
- ΔT – Period of time – from the Project beginning to the start of movement towards the SI Attractor.

The tuple of the system definition for ASI in the notations we have adopted verbally looks like this:

System "ASI (Cybernetics)" = < {blocks of CSs by strata}; their types; structures of the ASI CSs; Ecosystem and internal environment of ASI; CSs goals; Project team; Cybernetics; all stages of the Project >

The "ASI (Cybernetics)" system is a set of elements (CSs blocks) of various types, united and structured into a complex of control structures, that are functioning in the Ecosystem and in the internal environment of ASI with the goals (purposes) of management, observed by the Project Team at all its stages and described in the language of Cybernetics.

Special (cybernetic) properties and features of ASI as CS are described in chapter 25. Cybernetics (in the T&M Part), here we will analyze the system properties and features of CS:

- **Emergence** - all CSs (subsystems) of ASI work cooperatively and in concert to achieve common goals (possibly excluding simple regulation and homeostasis).
- **Hierarchy** - the static hierarchy of the ASI CSs structures is conditioned by (due to) the emergence of proactive target management.
- **Historicity** - development and management in ASI is continuous and "multi-pass" - all moves are recorded in the knowledge base and analyzed for further actions as experience.
- **Self-organization** - the main target (objective) function of ASI is self-improvement, that is, the CSs complex performs targeted management at the top level just for this. Although of course not every CS (subsystem) of ASI is capable of developing itself - some (perhaps most of them) will be improved with the help of other specialized subsystems.
- **Equifinality** - the main goal of ASI is to reach the SI Attractor, and equifinality has an existential meaning for ASI. Accordingly, management is aimed there.
- **Openness** - CSs (possibly excluding simple regulation) are open to the external or internal environment to receive information and issue control actions.
- **Non-equilibrium and non-linearity -** all complex upper-level control systems are the same.
- **Non-stationarity** and **dynamism** - It is possible to use the mechanisms of stochastization and "trembling" of parameters both in homeostasis systems and in other control systems.
- **Uniqueness, unpredictability and randomness -** It is possible to use positive feedback mechanisms in the target control systems and in their interaction with each other.
- **Adaptability - ability to adapt** - CSs with reflection, transformation, etc.
- **Variability - t**he ability to change behavior, structure and functional structure, while maintaining a holistic unity and basic properties. – adaptation and meta-adaptation plus homeostasis and meta-homeostasis.
- **Negentropy** - provided by external information and internal diversity (variety).
- **Purposefulness** - is provided by algorithms of target management.
- **The inconstancy** and **anisotropy** of internal spaces is still unclear for CSs.
- **Polystratic** - probably all CSs will be polystratic and will control the lower strata.

Setting the task for the PPR&D - all these system properties and features should be studied, analyzed, taken into account in the design documents.

35. Definition and system properties. Synergetics

Here we will formulate a complete definition of the ASI system as self-organizing, from the point of view of the Self-Organization SO-Theory or Synergetics.
First, we need to determine in which strata self-organization will occur:

- Material - no, development will be controlled by the upper strata
- Structural - neural networks will self-organize (or maybe not self?)
- Software - algorithms and data too
- Virtual - models and images too
- Intellectual - thoughts and concepts in the first place

SYSTEM - is a display of a finite set of objects with their properties and relations, isolated (selected) from the environment for a specific purpose, in the observer's language in a certain period of time. In symbolic form, this is a tuple:

$$S \overset{}{\underset{Def}{\equiv}} < A \,;\, Q_A \,;\, R \,;\, ENV \,;\, Z \,;\, N \,;\, L_N \,;\, \Delta T >$$

Where the components are particular definitions and general conditions, for ASI (Synergetics) these are:

- A - elements of the system - a set of (dynamic) artificial neurons, algorithms and databases, models and images, thoughts and concepts, people
- Q_A - elements properties – types and properties of dynamic elements.
- R - relations between elements - open non-linear dynamic structures.
- ENV - Environment - ASI Ecosystem, Mankind, Internet, Intertext, Infosphere, Noosphere.
- Z - goals (purposes) - self-improvement as self-organization.
- N - observer - the creators of the ASI - the Project team.
- L_N **-** the observer's language - Synergetics.
- ΔT – period of time – from the Project beginning to the start of movement towards the SI Attractor.

The tuple of the system definition for ASI in the notations we have adopted verbally looks like this:

System "ASI (Synergetics)" = < {dynamic elements by strata}; their types; dynamic ASI structures; Mankind and its spheres; self-improvement; Project team; Synergetics; all stages of the Project >

The "ASI (Synergetics)" system is a set of dynamic information elements of various types, united and structured into a complex of dynamic structures, which are functioning and self-organizing in Humanity in interaction with its spheres for the purpose of self-improvement, observed by the Project Team at all its stages and described in the language of Synergetics.

Special (synergistic) properties and features of ASI as a SO-system are described in chapter 26. Synergetics (T&M Part), here we will analyze the system properties and features of ASI from the point of view of the self-organization theory (Synergetics):

- **Emergence** - the target (objective) function of ASI is self-improvement, that is, purposeful self-organization, and, first of all, as an integral system.
- **Hierarchy** - dynamic hierarchy occurs in synergistic intellectual processes when order parameters arise and take control over the System behavior.
- **Historicity** - self-organization is a fundamentally historical process, the mechanisms of which require the continuous accumulation of changes and diversity (variety).
- **Self-organization** - from the point of view of Synergetics - is the main thing for ASI, in fact, its target (objective) function as a system as a whole is self-improvement.
- **Equifinality** - the system's own (by definition) desire for its Attractor, that is, self-organization.
- **Openness** - a permanent energy and information exchange with the environment on all strata.
- **Non-equilibrium and non-linearity -** It can be noted that already from the possibility of self-organization of ASI. Plus pumping energy and information.
- **Non-stationarity** and **dynamism** - and this is a necessary condition for the development and self-organization of ASI.
- **Uniqueness, unpredictability and randomness are** inextricably linked with the previous two and are also inherent in the mechanisms of positive feedback.
- **Adaptability - the ability to adapt** is important for stabilizing at new levels of organization.
- **Variability - the ability to change** behavior, structure and functional structure, while maintaining a holistic unity and basic properties. It is necessary for adaptation and development.
- **Negentropic** - provided by external energy, information and internal diversity - this is the basis for self-organization.
- **Purposefulness** - self-organization with a certain main goal – the SI Attractor.
- **Impermanence** and **anisotropy** - the complexity and diversity (variety) of interior spaces.
- **Polystratity** - We have identified four information strata for self-organization.

Setting the task for the PPR&D - all these system properties and features should be studied, analyzed, taken into account in the design documents.

36. Definition and system properties. Semiotics

Here we will formulate a complete definition of the ASI system as a semiotic (sign) system, from the point of view of the Signs Theory or Semiotics.

First, let us clarify - what signs are we talking about? From the point of view of Semiotics, all information processes inside (and outside) ASI are processes of semiosis (ch. 27. Semiotics in the T&M Part), and ASI itself and everything that is in it on information strata are semiotic (sign) systems, texts (hypertexts) and discourses (narratives). Moreover, the environment in which the ASI functions and develops is the Intertext and the Infosphere, which are semiotic supersystems too.

We determine in which strata semiosis occurs (representations and meta-representations):

- Material - contains extensionals and denotates and material carriers of signs
- Structural - happening
- Software - happening
- Virtual - happening
- Intellectual - happening

It should be noted that semiosis occurs simultaneously in three strata: signs proper and syntactic relations are located in the middle (although any sign has some material embodiment, but this does not apply to semiosis); intensionals, designates and semantics - in the upper; and extensionals, denotatates and pragmatics - in the lower stratum and in the external environment (or in any stratum in general, if we expand these concepts to non-material information objects). Thus, semiosis is a mechanism for establishing links between the strata of polystratic systems.

A more complex interpretation of semiosis in ASI implies the replacement of extensionals and denotatates of the external environment with their models in the internal space of ASI. But in this case, we have two processes of semiosis at once - internal with models (and between agents of the MAS - collective ASI?) and external with extensionals and denotatates in the external environment. External semiosis is the process of perception and communication (?) from the point of view.

SYSTEM - is a display of a finite set of objects with their properties and relations, isolated (selected) from the environment for a specific purpose, in the observer's language in a certain period of time. In symbolic form, this is a tuple:

$$S \underset{Def}{\equiv} < A \,;\, Q_A \,;\, R \,;\, ENV \,;\, Z \,;\, N \,;\, L_N \,;\, \Delta T >$$

Where the components are particular definitions and general conditions, for ASI (Semiotics) these are:

- A - elements of the system - signs: concepts, symbols, images, models.
- Q_A - properties of elements - types and properties of signs, intensionals, designates, intrasystem extensionals and denotates.
- R - relations between elements - sign structures and texts (hypertexts).
- ENV - Environment - Ecosystem and internal environment of ASI as text (hypertext) and discourse (narrative), Intertext and Infosphere. Any extensoinals and denotates are text in some sence too.
- Z - goals (purposes) - self-improvement as a discourse (narrative).

- N - observer - the creators of the ASI - the Project team.
- L_N - the observer's language - Semiotics and all languages used.
- ΔT – period of time – from the Project beginning to the start of movement towards the SI Attractor.

Thus, the tuple of the system definition for ASI (Semiotics) in the notations we have adopted verbally looks like this:

System "ASI (Semiotics)" = < {signs by strata}; their types; structures and texts of the ASI; internal and external environment of ASI as texts and extensionals; self-improvement discourse; Project team; Semiotics; all stages of the Project >

The "ASI (Semiotics)" system is a set of elements (signs) of various types, united and structured into a complex of sign structures and texts, functioning in the internal and external textual environment of ASI for the purpose of self-improvement discourse, observed by the Project Team at all its stages and described in languages of Semiotics.

Special (semiotic) properties and features of ASI as a sign system are described in ch. 27. Semiotics (in the T&M Part), here we will analyze the system properties and features of ASI from the point of view of the Sign Systems Theory (Semiotics):

- **Emergence** - ASI is a single (united) text and discourse that has internal coherence and generates a single meaning (narrative) in the upper intellectual stratum.
- **Hierarchy** - a complex multi-part text (hypertext) must be semantically hierarchical (and meta-hierarchical).
- **Historicity** – dynamic text and directed discourse is semantically historical
- **Self-organization** - in the semiotic sense, ASI as a text writes and rewrites itself!
- **Equifinality** - in the semiotic sense SI Attractor - also a text!
- **Openness** - a permanent exchange of information with the environment, (real world) extensionals and denotates are outside the ASI.
- **Non-equilibrium and non-linearity** - how is it in semiotics? It's not clear yet
- **Non-stationarity** and **dynamism** - similar
- **Uniqueness, unpredictability and randomness** - similar
- **Adaptability - Ability to adapt** - similar
- **Variability - The ability to change while remaining whole** and keeping the main thing - similar
- **Negentropy** - self-written text and self-sustaining discourse (narrative)
- **Purposefulness** – Purposeful Discourse
- **Impermanence** and **anisotropy** – semiosis up and down, internal and external…
- **Polystratic** – defined semiosis as fundamentally polystratic.

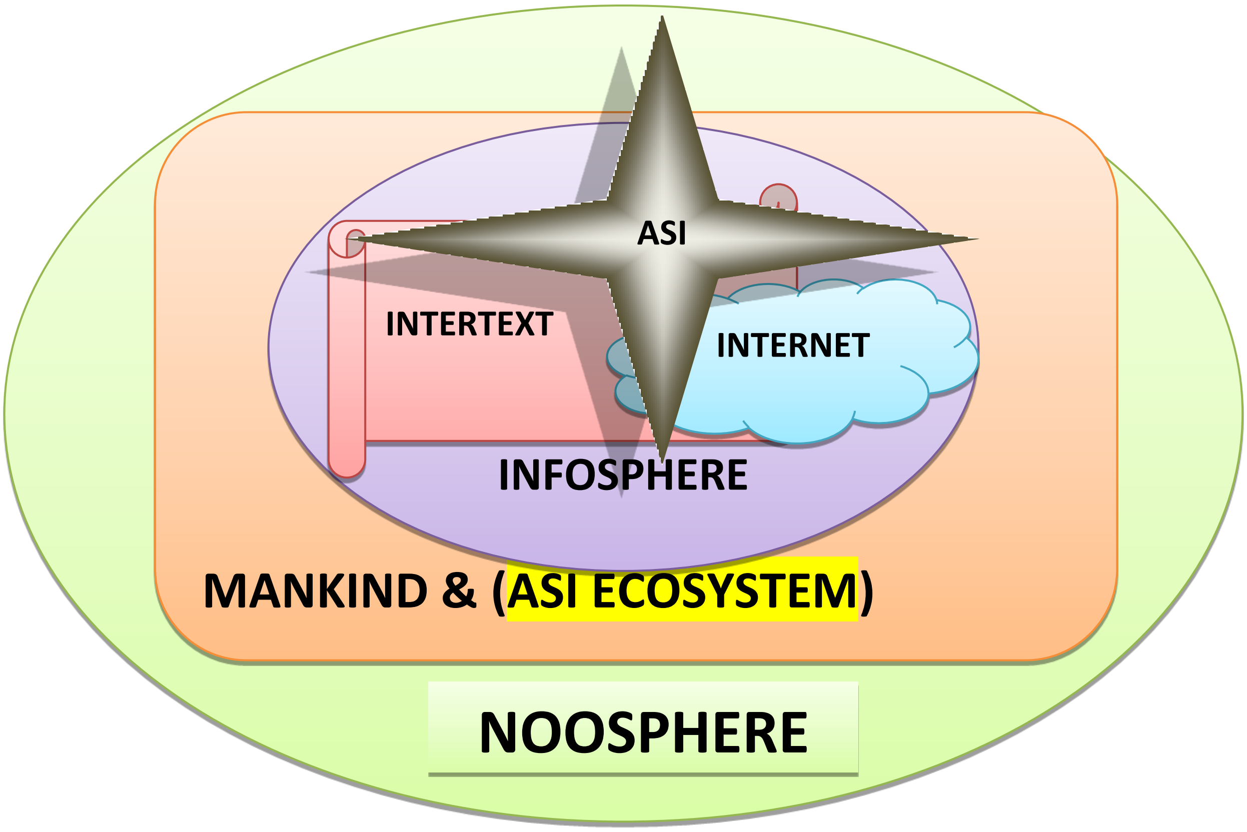
Setting the task for the PPR&D - all these system properties and features should be studied, analyzed, taken into account in the design documents.

37.United table by theories

DEFINITION	GTS / COGNITOLOGY	CYBERNETICS	SYNERGETICS	SEMIOTICS
Material elements	many units of main and auxiliary equipment (infrastructures)	Material embodiments of CSs blocks (where exist) ???	Material embodiments of info-(dynamic) structures???	Material embodiments of signs???
Structural elements	set of network nodes and artificial neurons	blocks of all control systems (ASI subsystems)	Sets of (dynamic?) artificial neurons	Signs - symbols
Program elements	set of algorithms and database	blocks of control algorithms	set of algorithms and database	signs: concepts, symbols, operators
Virtual elements	Sets of models and images	model blocks	Sets of models and images	signs: images, models
Elements of intelligence	Sets of thoughts and concepts maybe also people	blocks of intellectual processes, people	Sets of thoughts and concepts maybe also people	signs: concepts, symbols
Properties	description of types and characteristics of all elements			types and properties of signs, intensionals and designates, intrasystem extensionals and denotatates
Relationships	complex of all unifying structures	structures of control systems	open non-linear dynamic structures	sign structures and texts (hypertexts)
Environment	ASI Ecosystem, Humanity, Internet, Intertext, Infosphere, Noosphere	ASI Ecosystem, internal environment of ASI	ASI Ecosystem, Humanity, Internet, Intertext, Infosphere, Noosphere	internal environment of ASI as text (hypertext) and discourse (narrative), Ecosystem of ASI, Intertext and Infosphere
Goals (Purposes)	self-improvement, search and achievement of the SI Attractor	respectively, each CS has its own	Self-improvement as self-organization	self-improvement as a discourse (narrative)
Observer	creators of ASI - the Project team			
Language	natural languages, scientific and formal languages	Cybernetics	Cynergetics	Semiotics and all languages used
Period	from the Project beginning to the start of the movement towards the SI Attractor			

38. Environment and Supersystems

- ASI Ecosystem (Cyber-Physical) – Needs Elaboration (Appendix J)
- Internet = infrastructure + information + terminals + users
- Internet = digital (online) Infosphere = texts + infocommunication environment = Hypertexts
- Intertext = all texts created by Mankind (in the broadest sense) = online + offline = global context
- Infosphere = Intertext + information infrastructure + IT (in the broadest sense) + languages
- Infosphere = Internet + offline infrastructure and media + offline information (texts)
- Infosphere = Intertext + the entire infocommunication environment
- Humanity = people (and organizations) + artifacts + Infosphere
- Noosphere = Humanity + controlled Nature



So, let us try in ascending order - relations with ASI:

- **Internet**
  - Online texts (hypertexts)
  - Information environment (active)
  - Communication environment
  - NOT a supersystem
  - Receiving the information
  - Data storage
  - Communications
  - Collaborations
  - IT resources
  - Simulation environment

  - Virtual space (MetaVers)
  - Participation in development
  - Management (gradually)
  - Constraints - the life of the Internet

- **Intertext**
  - Includes all (online) texts of the Internet
  - Plus all texts are offline
  - Information environment (passive, global context)
  - Supersystem!!! - but not complete
  - Receiving the information
  - Participation in the formation and development
  - Management (gradually)
  - Goals (purposes) - integration, survival, expansion, progress of Mankind
  - Constraints - the life of the Intertext (Culture), Ethics?

- **Infosphere**
  - Includes Internet and Intertext
  - Plus offline communication environment
  - Information environment complete
  - Communication environment complete
  - Supersystem - but not complete!!!
  - Receiving the information
  - Data storage
  - Communications
  - Collaborations
  - IT resources - online and offline
  - Simulation environment
  - Virtual space (Virtual Reality VR, MetaVers MV)
  - Complemented space (Augmented Reality AR)
  - Participation in development
  - Management (gradually)
  - Goals (purposes) - integration, survival, expansion, progress of Mankind
  - Constraints - the life of the Infosphere, Ethics?

- **Humanity (Mankind)**
  - Includes Infosphere
  - Plus people (organizations) and artifacts
  - Supersystem!!! – complete?
  - Ecosystem (cyber-physical) for ASI
  - Information environment complete
  - Communication environment complete
  - Receiving the information
  - Data storage
  - Communications
  - Collaborations

- IT resources - online and offline
- All resources
- Simulation environment
- Operating environment
- Virtual space VR, MV
- Complemented space AR
- Real space
- Participation in development
- Management (gradually)
- Risks from people
- Goals (purposes) - survival, expansion, knowledge, progress of Mankind
- Constraints - Ethics?

- **Noosphere**
  - Includes Humanity
  - Plus Controlled Nature
  - Supersystem!!! – complete?
  - The material environment is incomplete (without the rest of Nature)
  - Information environment complete
  - Communication environment complete
  - Energy Exchange
  - Exchange of physical interactions
  - Receiving the information
  - Data storage
  - Communications
  - Collaborations
  - IT resources - online and offline
  - All resources
  - Environment for existence
  - Simulation environment
  - Operating environment
  - Virtual space VR, MV
  - Complemented space AR
  - Real space
  - Participation in development
  - Management (gradually)
  - Risks from people and Nature
  - Goals (purposes) - survival, expansion, knowledge, progress of Humanity, Evolution
  - Constraints - Ethics and Ecology?

Task Setting for PPR&D - all these environments and supersystems and their relationships with ASI should be studied, analyzed, taken into account in the project documents. Design an Ecosystem for ASI.

39. Element Analysis

So, we have five strata for analysis in the ASI system:

1. Material - iron (hardware) and electricity (infrastructures)
2. Structural - architecture and networks
3. Software - algorithms and data
4. Virtual - models and images
5. Intellectual - thoughts and concepts

In the GST definition of the system, elements are indicated - a set of units of main and auxiliary equipment (infrastructures), network nodes and artificial neurons, algorithms and databases, models and images, thoughts and concepts (intelligence elements), possibly (probably!) also people.

As far as possible at this stage, we will clarify the descriptions, characteristics and properties of the elements.

Material stratum - pieces of equipment (infrastructures):

- Basic equipment:
  - Server nodes of supercomputers - contain several types of processors (conventional central CPUs, graphics GPUs, tensor TPUs, neural network NPUs), operational and long-term non-volatile memory - in fact, full-fledged computers (rather even servers)
  - Server nodes of quantum computers - contain quantum processors (and memory?)
  - Remote ordinary and quantum computers included in a distributed network permanently or temporarily
- Auxiliary equipment:
  - Network equipment - connects server nodes and external elements
  - Energy equipment - power supply
  - Other technical equipment - control, cooling, etc.
- Terminal devices:
  - Sensors - controllers, modems, network cards, video cameras, scanners, radars, lidars, microphones, keyboards, touch panels, sensors, etc.
  - Actuators - controllers, monitors, displays, screens, projectors, acoustics, helmets and virtual reality glasses, printers, etc.
  - Controlled robots, drones and other individual devices
- People - concrete persons:
  - Participating in the implementation of the main ASI functions - in the Human-in-loop schemes HITL, Multi-agent systems MASs, collective, centauric and so on.
  - Maintenance, administrative and technical staff
  - Members of various collaborations

Structural stratum - structural/functional elements:

- Computer and network architecture (incl. quantum) – as in structure, not physically!
  - Server nodes of computers and external computers (servers)
  - Auxiliary equipment
  - Terminal devices

    - People's positions
- Neural networks (connectomes) - virtual, deployed in computers
    - Structural clusters of artificial neurons (incl. quantum)

Program stratum - directories (folders) or individual files:

- Operating System OS
    - OS functional blocks (probably) based on LINUX
    - I don't know yet - what will be there for quantum computers?
- Application programs
    - Separate applications for basic functions
- Utilities
    - Separate applications for other functions
- Data
    - Databases partitions
    - File Library Sections
    - External storage partitions
- People as special applications

Virtual stratum - models and images

- Models of ASI itself and individual agents in multi-agent systems
- Models of real objects and subjects of the external world
- Models of abstract (information) objects
- Template models for modeling
- Algorithms for creating and using models
- Sections of libraries and database of models and images

Intellectual stratum - ideas and thoughts

- Concepts - semantic units of dictionaries and thesauri (Semantics)
- Syntax rules of languages (Syntax)
- Knowledges (facts and text units) in knowledge bases KBs
- Thoughts - sentences in current discourse
- Ideas - saved thoughts
- Intelligent Algorithms

Setting the task for PPR&D - to work out the composition, types, characteristics and properties of elements. Determine the levels of exactly the elements and above the blocks and subsystems.

40. Structure Analysis

Here we try to describe the structures and individual subsystems by strata

Material stratum (infrasructures)

- LSICS supercomputers (possibly several different ones)
- Quantum computers
- Clusters of distant computers and structurally separate
- Terminal blocks and structurally separate
- Auxiliary systems
- Departments in the organization
- Groups and Individuals in Centauric Systems and Collaborations

Structural stratum

- Network structure
- Neural networks (connectomes)
- MASs structures
- Centauric systems with humans
- Organizational structure of the organization
- Collaboration structure

Program stratum

- OS operating systems
- Software Libraries
- Clusters of neural network algorithms
- Databases DBs
- Groups of agents (incl. people) as apps

Virtual stratum

- MetaVerses
- Internal spaces ISs
- Mental maps MMs
- Algorithm libraries for them
- Libraries (catalogues) of spaces and maps

Intellectual stratum

- Knowledge Bases KBs and Thesauri (Semantic graphs, metagraphs etc.)
- Languages, metalanguages, hyperlanguages, etc.?
- Intelligent subsystems (incl. agents in MAS)?
- Subsystems of Consciousness (such as CTM, AMI and/or others)

The task statement for PPR&D is to work out and draw all structures on all strata with connections and hierarchy.

41. Function Analysis

We will also briefly list by strata and structures

Material stratum

- Supercomputers - physically: digital operations and digital memory, deployment of virtual neural networks, other computer functions
- Quantum computers - physically: quantum computing and other operations
- Distant computers - physically: distributed computing and memory
- Terminal devices - physical sensors and actuators - exchange of information with the environment, informational and physical effects on the environment.
- Auxiliary systems - energy and technical support and service
- Organization (groups and teams) - maintenance of equipment (infrastructures) by people, work and personnel management
- People in collaborations & Centauric MASs- joint execution of works and functions

Structural stratum

- Network structure - information and energy interaction between elements and subsystems, distribution of functions and flows (pipelines) of energy and information.
- Neural networks (connectomes) - interaction between neurons and clusters
- Organizational structure of the organization (enterprise) - the interaction of employees and departments, management
- Collaboration & MAS structure - interaction and management in collaborations and MASs

Program stratum

- Operating systems OSs - support of the internal operating environment for all application programs and algorithms, homeostasis, interaction with lower strata
- Application programs - performing all computer functions
- Neural network algorithms - performance of standard intellectual functions, support of the internal environment for self-organization, learning and development of intelligence
- Databases - storing information in the form of data
- Groups of people as applications - functional organization and collaborations and MASs

Virtual stratum

- Internal spaces IS (hyperspaces) - support for the internal environment for the placement and interaction of mental maps and subspaces with models
- Mental maps MM - maps/subspaces for placement and interaction of models of objects and subjects, real, physical and abstract.
- Algorithm libraries - storing and providing them for use
- Catalogs of spaces and maps - metamaps of created internal spaces and maps

Intellectual stratum

- Knowledge bases KB – creation, storage and provision of knowledge

- Thesauri - systematization of knowledge, creation, development and support of semantic networks (knowledge graphs KG)
- Languages - internal and external communications and information handling at an intellectual level, support for semiosis
- Intelligence functions and algorithms, including (possibly):
    - Guided (controlled) Perception
    - Search, gathering and analysis of information
    - Goal setting and planning
    - Forecasting and prediction
    - Search and decision making
    - Action management
    - Abstract thinking - operations with concepts and images
    - Logical thinking (logic) - reasoning
    - Communication using sign systems (for example, language)
    - Managed memory
    - Cognition, learning and self-learning
    - Professional activity
    - Self-awareness and reflection
    - Values & Ethics
    - Empathy
    - Motivation
    - Aesthetics
    - Creation
    - Imagination & Dreams
    - Games & Entertainment
    - Humor
    - Other
- Special functions of AI and BM
- Functions of the subsystem/model of Consciousness (such as CTM, AMI and/or others)
- Especially – "indicator properties" of consciousness
- Continuous episodic memory
- Mental Time Travel MTT
- Agents - creation, development and use of internal intelligent agents and multi-agent systems, including with people
- Teams and collaborations - with humans and other AIs
- Collective ASI - people + AI systems

The task setting for the PPR&D is to work out all the functions on all strata and in blocks.

What to lay down previously (preliminarily) - before initiation, and what late - during training, self-organization, and self-training.

42. United table strata-elements-structures-functions

	ELEMENTS	STRUCTURES	FUNCTIONS
MATERIAL STRATUM (infrastructures)	• Basic equipment • Auxiliary equipment • Terminal devices • People – persons (individuals)	• **(infrastructures)** • LSICS supercomputers (possibly several different ones) • Quantum computers • Clusters of distant computers and structurally separate • Terminal blocks and structurally separate • Auxiliary systems • Departments in the organization • Groups and Individuals in Centauric MASs and Collaborations	• Supercomputers - physically: digital operations and digital memory, deployment of virtual neural networks, other computer functions • Quantum computers - physically: quantum computing and other operations • Distant computers - physically: distributed computing and memory • Terminal devices - physical sensors and actuators - exchange of information with the environment, informational and physical effects on the environment. • Auxiliary systems - energy and technical support and service • Organization (groups and teams) - maintenance of equipment by people, work and personnel management • People in collaborations and MASs – joint work
STRUCTURAL STRATUM	• Computer and network architecture • Neural networks - virtual, deployed in computers	• Network structure • Neural networks (connectomes) • Centauric systems with people • Organizational structure of the organization • Collaboration structure • MASs structure	• Network structure - information and energy interaction between elements and subsystems, distribution of functions and flows (pipelines) of energy and information. • Neural networks - interaction between neurons and clusters • Organizational structure of the organization - the interaction of employees and departments, management • Collaboration structure - interaction and management in collaborations and MASs

	ELEMENTS	STRUCTURES	FUNCTIONS
PROGRAM STRATUM	• OS operating systems • Application programs • Utilities • Data • People as special applications	• OS operating systems • Software Libraries • Clusters of neural network algorithms • Database DB • Groups of people as apps	• Operating systems OS - support of the internal operating environment for all application programs and algorithms, homeostasis, interaction with lower strata • Application programs - performing all computer functions • Neural network algorithms - performance of standard intellectual functions, support of the internal environment for self-organization, learning and development of intelligence • Databases - storing information in the form of data • Groups of people as applications - functional organization and collaborations and MASs
VIRTUAL STRATUM	• Models of the ASI itself • Agent Models in MASs • Models of real objects and subjects of the external world • Models of abstract (information) objects • Template models for modeling • Algorithms for creating and using models • Sections of libraries and database of models and images	• MetaVerses MV • Internal spaces IS • Mental maps MM • Algorithm libraries for them • Libraries (catalogues) of spaces and maps	• Internal spaces IS (hyperspaces) - support for the internal environment for the placement and interaction of mental maps and subspaces with models, MetaVers funtions • Mental maps MM - maps/subspaces for placement and interaction of models of objects and subjects, real, physical and abstract. • Algorithm libraries - storing and providing them for use • Catalogs of spaces and maps - metamaps of created internal spaces and maps

	ELEMENTS	STRUCTURES	FUNCTIONS
INTELLECTUAL STRATUM	• Concepts - semantic units of thesauri (Semantics) • Syntax rules of languages (Syntax) • Knowledge in knowledge bases • Thoughts - sentences in current discourse • Ideas - saved thoughts • Intelligent Algorithms	• Knowledge Bases KB and Thesauri • Semantic graphs & metagraphs • Languages, metalanguages, hyperlanguages? • Intelligent subsystems? • Subsystem of Consciousness (CTM type, AMI or other)	• Knowledge bases KB – creation, storage and provision of knowledge • Thesauri - systematization of knowledge, creation, development and support of semantic networks (KG) • Languages - internal and external communications and information handling at an intellectual level, support for semiosis • Intelligent functions and algorithms, including (possibly) processes (the above) • Special functions of AI and BM • Especially – "indicator properties" of consciousness • Continuous episodic memory • Mental Time Travel MTT • Functions of the subsystem/model of Consciousness (such as CTM, AMI and/or others) • Agents - creation, development and use of internal intelligent agents and multi-agent systems MAS, including with people • Teams and collaborations - with humans and other AIs • Collective ASI - people + AI systems

43. Input-Output Analysis

Material stratum

MODALITIES	INPUTS	OUTPUTS
Mechanics, sound EM waves, light Thermal energy Electricity Chemistry	Getting energy and any impact from the external environment	Transfer of energy and any impact to the external environment
TERMINAL DEVICES	Sensors - controllers, modems, network cards, video cameras, scanners, radars, lidars, microphones, keyboards, touch panels, sensors, etc.	Actuators - controllers, monitors, displays, screens, projectors, acoustics, helmets and virtual reality glasses, printers, etc.
ROBOTS	Controlled robots, drones and other individual devices	
PEOPLE	People with whom ASI interacts	
ENVIRONMENT	PHYSICAL SPACE AND OBJECTS, NOOSPHERE	

Structural stratum

Information - receiving and transmitting signals over networks
ENVIRONMENT - information and power networks, INTERNET

Program stratum

Data - receiving and transferring files, commands and requests, metadata
ENVIRONMENT - INTERNET, INFOSPHERE

Virtual stratum

Images - receiving and transmitting signs
ENVIRONMENT - INTERTEXT, INFOSPHERE, HUMANITY, NOOSPHERE, ASI ECOSYSTEM

Intellectual stratum

Texts - receiving and transmitting concepts (meanings)
ENVIRONMENT - INTERTEXT, INFOSPHERE, HUMANITY, NOOSPHERE, ASI ECOSYSTEM

PEAS – definition of the ASI agent space:

- Performance measurement (objective function) - self-improvement
- Environment- Intertext, Infosphere, Humanity, Noosphere, ASI Ecosystem
- Actuators - terminal devices, robots, people
- Sensors - terminal devices, robots, people

The task statement for the PPR&D is to work out all the inputs and outputs on all strata and between

44. Processes Analysis

By processes, in contrast to just functions, we will understand regular ordered actions to transform certain input resources (inputs) into results (outputs). In this case, the process may contain several different functions used (sequently and/or parallel) for this transformation.

Through polystratic processes

- **Perception** - continuous receipt of information from the outside world (environment) and inside?
- **Self-consciousness -** awareness, attention, continuous episodic memory, MTT, Active inference? etc.
- **Communication** - two-way exchange of information with an external subject and inside in MASs?
- **Activity** - controlled purposeful (target-directed) actions to solve specific problems
- **Learning** - an activity with the aim of acquiring/improving any abilities

Description: - tables by strata. While exemplary

- Input source
- Input
- Input Format
- Functions (by performers)?
- Output
- Output Format
- Output Receiver (Destination)

PERCEPTION - for now, rather as an example of an end-to-end bottom-up process by strata (not by blocks - the block diagram will be more detailed, see, for example, in the work mentioned below)

The table was compiled based on the description of the human perception apparatus in the author's early paper in book [Новиков (2022)], however, in a different section of the strata and with the addition of spaces of virtual models.

STRATA	SOURCE	INPUT	FORMAT	FUNCTIONS (by)	OUTPUT	FORMAT	DESTINATION
Material	Sensors - different	EM and other fluctuations	Different modalities	Quantization discretization (processors)	Electric signals	Discrete quantized	Neural networks
Structure	Processors	Electric signals	Discrete quantized	Filtration (neural networks)	Data	Processed	Algorithms
Program	Neural networks	Data	Processed	Glossy analysis (algorithms)	Glosses	Identified	Virtual spaces
Virtual	Algorithms	Glosses	Identified	Synthesis & update of models (virtual spaces)	Models of extenal world	Updated	Intelligence
Intellect	Virtual spaces	Models of extenal world	Updated	Understanding (intellect)	Thoughts	Text	Intelligence function

SELF-CONSCIOUSNESS

Continuous episodic memory allows the subject (intelligent agent) to perceive himself in the environment (space) and time, and most likely, this is one of the foundations of Consciousness

Mental time travel MTT expands the functionality of episodic memory to the ability not only to feel here and now, but also to project into any alternative past and simulated imaginary future. In fact, it is the basis of the higher functions of Consciousness.

Active Inference - self-evidence, continuous beliefs updating and propagation, model evidence optimization, sharing of narratives, goals and models between agents in MAS and collective ASI, etc.

Different functions of Consiousness (from many models) – awareness, attention, etc. Inputs-outputs, stratification and set of functions can be very different and non-obvious, it may be necessary to classify.

COMMUNICATIONS

Three sub-processes (functions) = controlled perception of partner text + creation (generation) of own text + transfer of own text to partner

1. The perception of a partner text is one of the perception process types
2. Create your own text - various intelligence functions can be used
3. The transmission of a text to an external partner is a descending process from the intellectual to the material stratum, in general, the opposite process of perception.

The inputs and outputs of the communication process can be different depending on the tasks, content, format, partners, context, etc.

ACTIVITY

- Operating - with an impact on the external environment, including in the material world
- Intellectual - only information, maybe even everything can be inside the ASI itself.

Activity management can be targeted or by deviation (regulation), general cybernetic algorithms are given in [Новиков (2012)].

Inputs-outputs, stratification and set of functions can be very different, it may be necessary to classify.

LEARNING

- Is it possible to consider learning (training) as a kind of activity?
- The input can be initial (before learning) testing (as a trigger) and the necessary resources.
- The output is new/improved abilities (inside) and (post learning) testing (external and internal).
- It can also be entirely inside - on models of internal spaces.
- Also important is meta-learning, that is, learning to learn.

There are many models of these processes, including for AI - this is generally a separate big topic.

The task setting for the PPR&D is to work out the main processes at all strata and by functions, possibly in variants and with a classification. Select notation. What to lay down preliminarily - before initiation, and what later - during training, self-organization, and self-training?

45. Phase space behavior Analysis

Initially, it was designated as follows - Analysis of behavior, history and dynamics in the phase space. However, our designed system simply does not have a history and observed dynamics yet, and we can try to describe the expected behavior only.

The phase space is a multidimensional space of general (key) system parameters. The system state is a vector (dimensioned by the number of parameters) in the phase space with the coordinates of the parameters values at each moment of time.
Behavior is a change in these parameters, that is, a trajectory of movement in the phase space.

As a first approximation, the key parameters are:

- Level of complexity (perfection, organization) - the main target parameter.
- Level of readiness (completeness) for initiation
- Level of readiness for the start of movement to the SI Attractor
- Scope of knowledge
- Computing power
- Number of involved neural networks parameters (synapses)
- Parallelism of processes (functions)
- Intelligence level (IQ analogue)
- Tests for various intellectual functions (and consciousness, incl. indicator properties))
- Accuracy of smart functions and models (% errors)
- Efficiency = results/resources
- Number of individual computers in distributed networks and collaborations
- Number of people in centauric MASs and their effectiveness
- Number of employees of related organizations and their effectiveness
- Number of people in collaborations and their effectiveness
- Energy consumption and energy efficiency
- Financial and economic parameters of the Project
- Parameters for the directions - GST, Cybernetics, Synergetics, Semiotics, Cognitology
- Options to provide system properties
- Parameters by stratum and Parameters by functions and processes
- What else?

The parameters will need to be classified and summarized in tables (matrixes) and subspaces.
To form a phase space of an acceptable dimension, it may be necessary to reduce the number of parameters or reduce them to a few of the most important ones. Maybe apply synergetic tools and highlight the order parameters. The optimal dimension number must be justified.

Behavior

It is clear that the target and performance parameters should increase with the development of ASI, possibly in stages. There may also be patterns of behavior for specific tasks.

The problem statement for the PPR&D is to form a phase space with the optimal number of key parameters, describe them, and analyze the behavior and its patterns.

46. Goals and Objective functions

Here we do not confuse the Project goals with the goals of the ASI system itself. The Project must first create an ASI and bring it to initiation - that is, to the beginning of autonomous self-government and self-improvement, and only then do the actual goals for the ASI appear, but only until the start of movement towards the Attractor - then the ASI itself sets the goals. The Project goals (objectives) at the stage before the ASI initiation are defined and decomposed in sufficient detail in the STRATEGY Part, and the ASI Priorities at the stage (trajectory) of movement towards the Attractor are preliminarily indicated there.

Goals (Purposes) of the ASI System - Self-improvement, search and achievement of the ASI Attractor.

- **Self-improvement** - improvement of intellectual skills, as a specific goal - to the level necessary to detect and start moving towards the SI Attractor
- **SI Attractor** - a "perfect" ASI, capable and striving to fulfill its mission (the mission of ALL Strong Intellects SI) - to lead the transition of Humanity to Posthumanity

Goals (Purposes) from the external Environment (higher level Supersystems) - survival, expansion, knowledge, progress of Humanity, Evolution

- **Survival** - the survival of Mankind (Humanity) as a Civilization (not necessarily as a biological species Homo Sapiens Sapiens only)
- **Expansion** – expansion of the Mankind habitat (global areal - Noosphere), Space expansion.
- **Cognition** – collection and improvement the knowledge of Nature and the Universe.
- **Progress** - improvement, increase in organized complexity and decrease in entropy, complex progress - scientific and technical progress STP, social, cultural, etc.
- **Evolution** - the evolution of the Universe in the context of the Universal History.

Note that the goals from the Environment will become ASI goals already on the trajectory to the Attractor, and before that - they will be rather (mandatory) restrictions (constraints) and guidelines in interaction with the Environment.

Objective (target) functions - continuous (permanent) self-improvement

- Increasing and improving abilities (skills, powers)
- Increasing and improving knowledge
- Improving the relevant parameters - we need to select the key (see previous chapter)
- What else?

Setting the task for the PPR&D is to form a system of goals and objective functions, formulate them, and determine the parameters and their target values.

47. Goals Decomposition

Very preliminary, yet more like thoughts and sketches.

By strata

- Material
  - The complexity and power of the supercomputer system (including controlled external systems) is two to three orders of magnitude higher than the human brain
  - The amount of available memory - comparable to the entire Internet (including external controlled servers and computers)
  - Complete quantum computers (networks)
  - Terminal devices - full (sufficient) control of the (near, required) environment
  - Collective ASI - people + AI systems - completeness and sufficiency
- Structural
  - Neurons in neural networks - an order of magnitude more than in the brain - Trillion!!!
  - Active parameters of neural networks (that is, connections) - A thousand trillion!!!
  - Collective ASI?
  - Control (full, sufficient) of the Internet
- Software
  - A complete package of all applications for any computer functions
  - Full control of these applications plus the ability to autonomously improve them and develop any new ones
  - Own (self-developed) internal programming languages?
  - Programs for quantum computers
  - Controlled databases - all that are and may be needed
- Virtual
  - A complete functional models of all external environments with the required detail
  - Models of the ASI itself, the internal environment and MAS agents (incl. people)
  - Control of external virtual worlds and MetaVerses
- Intellectual
  - Full functionality of individual human Intelligence (and consciousness)
  - Full functionality of Group Intelligence
  - The power of functions is 2-3 orders of magnitude higher than human
  - ALL knowledge of the Infosphere is available and functional

By key intellectual processes

- Perception - sufficient control of the environment in real time and inside?
- Self-consciousness – episodic memory continuity and managed MTT, Active Inference, etc.
- Communication - free effective communication with any subject and inside?
- Activities - 100% effective implementation of all (any) tasks
- Learning - 100% effective autonomous training for what you need

The problem statement for the PPR&D is to form a decomposed hierarchical system of goals and objective functions, formulate them, and determine the parameters and their target values.

48. Data issues

General considerations about data for the ASI systems

- **What is the data for**
    - Learning
    - Cognition
    - Activity
    - Self-improvement
    - Self-organized collectivization
- **What data - by strata**
    - Software – data and metadata in the form of multidimensional arrays (tensors*spectra*hypercomplex*probability distributions) and quantum? Also logical constants - fuzzy and temporary? Files = texts!!! Graphs?
    - Virtual - signs, images, models. Also texts (hypertexts).
    - Intellectual - concepts and texts (hypertexts). Facts and knowledge.
- **What is already there**
    - DBs (corpora) for AI training - an overview is below in chapter 50. Data for BMs
    - Internet - Wikipedia, dictionaries, libraries, social networks, different websites, etc.
    - Specialized databases for various topics and activities, sciences, etc.
- **Where and how to pick them up**
    - Review of ready databases used for BMs is below in chapter 50. Data for BMs
    - Form your own databases with information from the Internet and available sources
    - It is necessary to teach ASI to make a database from any available information!!!
    - Well, make Knowledge Bases KB and Thesauri from the database
- **Where and how to store**
    - On their servers as part of the ASI system
    - In external servers and computers in organized distributed systems
    - In general, in any external storage - servers, computers, clouds
    - In distributed databases organized on the upper floors
- **What and how to do with them**
    - Create - organize, form, structure, fill
    - Develop - expand, deepen, refine, improve, update
    - Use for self-development and any current tasks
- **What is needed for this**
    - Tasks and Requirements
    - Technologies and Resources
    - Algorithms and Models
    - Iron (hardware, infrastructure) and People

Setting the task for the PPR&D - Work through all issues and solutions

49. About data from other sections

From the chapter 30. Mathematics (last three points added):

ASI must operate with data (values, variables) that are/have mathematical properties at the same time and belong to the relevant sections of mathematics and related disciplines:

- **Tensors** - in our real three-dimensional world, in general, all quantities are tensors of at least the third order - Tensor Analysis
- **Spectra** in the frequency domain - in general, all time-varying quantities have (can be decomposed) a frequency spectrum - Harmonic Analysis
- **Complex** (Hypercomplex) Numbers - Complex and Hypercomplex Analysis
- **Probability** Distribution of a Value – Probability Theory and Mathematical Statistics
- **Logical** constants and variables - at least second order, fuzzy and temporary - Discrete Mathematics
- **Graphs** - represented as various graphs (meta-, hyper-, factor-, etc.) – Graphs Theory
- **Quantum** Constants and Variables - Quantum computing

From the chapter 29. Artificial Intelligence - Information for the development of ASI

- **Big Data -** in general, everything that is possible + the entire context, including (and mostly) raw data and (of cause) corpora for AI machine learning.
- (Free) **Open sources** - searching and discovering, creating your own
- From the total number of open databases to the **specific models** - including sharing blocks and subsystems for testing, refinement and development in network **collaborations and crowdsourcing**
- **Post-structuralism and Hermeneutics** - knowledge as a text (hypertext) in the fullness of its **context**, external and internal **relations**, including **history**, the **personality** of the author and even the agent-"reader" in the ASI (MAS) system
- **Hypertext** (Superhypertext) - connection of all information (so far on the Internet) into a united database system and Knowledge Base for ASI

From the chapter 38. Environment and Supersystems - Infosphere as an information environment with data for ASI

- Includes Internet and Intertext
- Plus offline communication environment
- Information environment complete
- Communication environment complete
- Receiving the information
- Data storage
- Communications
- Collaborations
- IT resources - online and offline

- Simulator
- Virtual space (VR and MetaVers MV)
- Complemented space (Augmented reality AR)
- Participation in development
- Management (gradually)

Also, the cyber-physical ecosystem of the collective ASI - to work out

From the chapter 28. Cognitive Science (Cognitology)

Knowledge differs from a simple data in a number of some essential properties:

- the unit of information processed is a fact
- internal interpretability
- activity
- connectedness
- structuredness
- semantic metric
- convertibility of representations

Fact is a record of data, resulting semantics:

- Name
- Meaning (Value);
- the degree of confidence in the validity of the value;
- many connections
- set of allowed functions

The knowledge base (KB) is a database containing the actual knowledge and inference rules in a certain subject area. In self-learning systems, knowledge base also contains information that is the result of previous learning and activities - that is, experience.

Semantic web (network) - a semantically structured knowledge base, an information model of the subject area, has the form of a directed graph. The vertices (nodes) of the graph correspond to the objects of the subject area, and the arcs (edges) define the relationship between them. Objects can be concepts, events, facts, properties, processes, in general - any knowledge and its elements. Edges are predicates and functions in first-order logic.

A semantic (kowledge) graph (KG) is a formalization of a semantic network, or just a synonym

Thesaurus = the Knowledge Base in a specific subject area/domain (can be represented both as a dictionary with semantic links and as a semantic graph)

For ASI, a universal Ontology is needed - a general ontology. And further from [Russell & Norvig (2021)] - a summary of the development of ontologies in the field of AI:

- Ontological engineering
- Upper Ontology, Category, subcategory, inheritance, taxonomy, (de)composition
- Measure, unit, natural kind, mass & count nouns, in- & extrinsic properties, events
- Mental objects, modal logic, temporal logic, description logic
- Circumscription & default logic, truth maintenance
- Qualitative physics, spatial reasoning, psychological reasoning

Statement of the problem for PPR&D

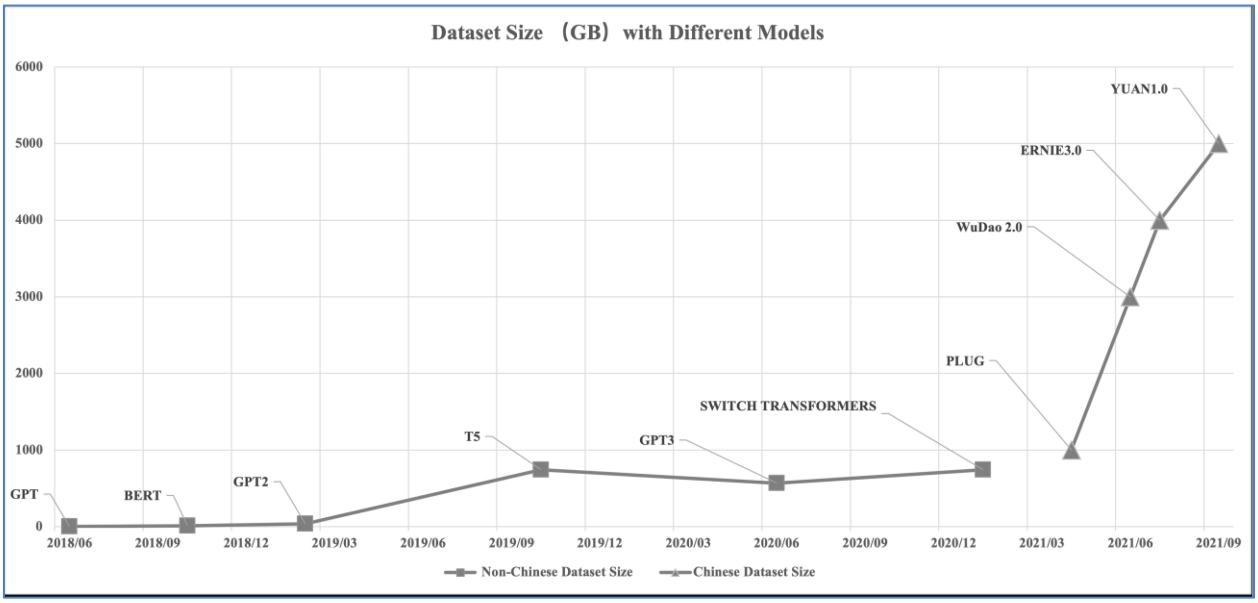
- Work out the mathematical requirements for the data
- Work on Poststructuralism and Hermeneutics
- Work out Hypertext
- Develop the Infosphere and Ecosystem of Collective ASI
- Work out Ontologies
- Work out Knowledge bases KB, semantic graphs

50. Data for Big Models

Based on [RM for BM (2022)]

Big Models BMs (see Appendix H) = MegaData + MegaComputers + Intelligent algorithms

- Big data Driven
- Multi-tasks Adaptive
- Few-shot (Zero-shot) Learning



<u>**Datasets (corpora) for BMs training – now (march 2022) used (Size and for witch BMs):**</u>

- English Wikipedia - 19.13GB - BERT, XLNet, GPT3
- BookCorpus2 - 9.45GB - BERT, XLNet, RoBERTa, GPT3
- RealNews - 120GB - Grover
- OpenWebText2((OWT2)) - 125.54GB - GPT2/3, RoBERTa
- PubMed Central - 180.55GB - GPT-neo, BioBERT
- ArXiv - 112.42GB - GPT-neo, WuDao
- C4 - 750GB - T5
- Wiki-40B - 4GB - Transformer-XL
- CLUECorpus2020 - 100GB - RoBERTa-large-clue
- The-Pile - 1254.20GB - GPT-neo, WuDao
- CC100 - 2.5TB - XLM-R
- multilingual C4(mC4) - 26TB - mT5
- Conceptual Captions(CC) - 3.3M image-text pair - VL-BERT
- LAION-400 - 400M image-text pair - CLIP, DALL-E
- WuDaoCorpora 650M image-text pair + 5TB - CPM-2, WuDao

Working with Data for BMs

- **Corpora Construction**
- **Generate Database DB on Big Model and Knowledge Graph KG**
- **Multimodal Fusion**
- **Knowledge Graphs and Knowledge Integration/Fusion**
    - Experts annotated knowledge graphs
    - Wiki-Based knowledge graphs
    - Knowledge graphs extracted from unstructured texts
- **Knowledge Graph Completion and Integration**
    - Link Prediction
    - Entity Alignment
    - Entity Matching
    - Entity Linking (i.e. Wikification)
- **Big Model-based Knowledge Acquisition**
    - Big Model as Booster for Knowledge Acquisition
        - Encoder and Fine-tuning
        - Parameter-less Tuning
        - Machine Reading Comprehension & QA Paradigm
    - Big Model as Resource for Knowledge Acquisition
        - Big Models for Data Augmentation
        - Big Models are Knowledge Bases
- **Knowledge-enhanced Big Models**
    - Commonsense and Domain-specific knowledge
    - Knowledge Graphs as Side Information
        - integrate knowledge graph representations
        - betteralignment with more informative contexts
        - learn native entity representations
        - external knowledge memory
        - use the knowledge graphs to guide or improve the challenge of language pre-training
        - improve language generation with knowledge graphs
    - Learning Knowledge Graph Abilities

Perspectives

- **Learning the Ability Rather Than Information of Knowledge Graphs**
    - the multi-hop symbolic reasoning to acquire new knowledge
    - the hierarchical conceptual abstraction
    - the structural information compression
    - and the condensation of human consensuses
    - the meta-knowledge of operating over knowledge graphs
    - the external knowledge graph memory

- **Introducing More Genres of Information in Knowledge Graphs**
    - cross- and multimodal big models
    - new qualifiers and attributes

Statement of the PPR&D problem – work out issues of data and knowledge in the BM paradigm

51. Necessity and Sufficiency

One thing is necessary for another, if the other cannot appear without the appearance of the first

- **Theory**
    - Systems Theory GST - full system synthesis
    - Cybernetics - all control subsystems CSs
    - Synergetics - conditions for self-organization
    - Cognitology - intellectual functions
    - Consciousness - models of consciousness

- **Methodology**
    - Interdisciplinarity
    - Complementarity
    - Key, frontier and promise AI technologies
    - Imitation of human intelligence development?

- **System**
    - Stratification - all strata and by strata
    - Material – key specification parameters no less than the human brain
    - Structural - neural networks
    - Software – primarily embedded structures and algorithms (which are necessary)
    - Virtual - internal MetaVers with models
    - Intellectual - languages, algorithms
    - Energy supply
    - Collectivity (multi-agency - MAS)
    - Collective ASI Ecosystem

- **Data**
    - Information Support
    - Data and knowledge corpora for education and training

IN GENERAL – COMPLIANCE WITH ALL CONDITIONS!!!

One thing is sufficient for another if the appearance of the first ensures the appearance of the other.

- **Theory**
  - Systems Theory GST - system redundancy (superfluity, abundance in every sense)
  - Cybernetics - excessive variety at all levels. Models of ASI and Environments in CSs!
  - Synergetics - Non-linearity, complexity, non-stationarity, etc.
  - Cognitology- ALL cognitive science
  - Consciousness - ALL models of consciousness and indicator properties

- **Methodology**
  - ALL AI methodology
  - Higher order algorithms and metaalgorithms
  - Models, metamodels, hypermodels, etc.
  - Reinvestment of results (in one sense or another or in all)

- **System**
  - Material - parameters are much higher than the brain and the possibility of growth, supercomputers and quantum computers
  - Structural - excess Communication
  - Software - an excess of structures and algorithms
  - Virtual - several MetaVerses? Models of everything and everywhere, evidence, etc.
  - Excess energy

- **Data**
  - Too much information

IN GENERAL – EXCESSIVE (SUPER/OVER ABUNDANCE) IN ALL!!!

52. Summary of Conceptual Model

System Definitions by Theories

GST & Cognitology: The "Artificial Super Intelligence ASI" system is a set of material and informational elements of various types, united and structured into a complex of special structures, that are functioning in Humanity in interaction with its spheres in order to achieve the SI Attractor, observed/controlled by the Project Team at all stages, and described in natural and formal languages.

Cybernetics: The “ASI” system is a set of elements (CSs blocks) of various types, united and structured into a complex of control structures, that are functioning in the Ecosystem and in the internal environment of ASI with the goals (purposes) of management, observed by the Project Team at all its stages and described in the language of Cybernetics.

Synergetics: The "ASI" system is a set of dynamic information elements of various types, united and structured into a complex of dynamic structures, which are functioning and self-organizing in Humanity in interaction with its spheres for the purpose of self-improvement, observed by the Project Team at all its stages and described in the language of Synergetics.

Semiotics: The "ASI" system is a set of elements (signs) of various types, united and structured into a complex of sign structures and texts, functioning in the internal and external textual environment of ASI for the purpose of self-improvement discourse, observed by the Project Team at all its stages and described in languages of Semiotics.

System Properties by Theories are identified and described

- Emergence
- Hierarchy
- Historicity
- Self-organization
- Equifinality
- Openness
- Non-equilibrium and non-linearity –w/o semiotics? It's not clear yet
- Non-stationarity and dynamism - w/o semiotics? It's not clear yet
- Uniqueness, unpredictability and randomness - w/o semiotics? It's not clear yet
- Adaptability - w/o semiotics? It's not clear yet
- Variability - w/o semiotics? It's not clear yet
- Negentropy
- Purposefulness
- Impermanence and anisotropy - w/o cybernetics? It's not clear yet
- Polystratity

Environment and Supersystems

- ASI Ecosystem (Cyber-Physical) – Needs Elaboration
- Internet = infrastructure + information + terminals + users
- Internet = digital (online) Infosphere = texts + infocommunication environment = Hypertexts
- Intertext = all texts created by Mankind (in the broadest sense) = online + offline = global context
- Infosphere = Intertext + information infrastructure + IT (in the broadest sense) + languages
- Infosphere = Internet + offline infrastructure and media + offline information (texts)
- Infosphere = Intertext + the entire infocommunication environment
- Humanity = people (and organizations) + artifacts + Infosphere
- Noosphere = Humanity + controlled Nature

Elements and Structures by strata

	ELEMENTS	STRUCTURES
MATERIAL STRATUM (infrastructures)	• Basic equipment (hardware) • Auxiliary equipment • Terminal devices • People – persons (individuals)	• **(infrastructures)** • LSICS supercomputers (possibly several different ones) • Quantum computers • Clusters of distant computers and structurally separate • Terminal blocks and structurally separate • Auxiliary systems • Departments in the organization • People in Centauric MASs and Collaborations
STRUCTURAL STRATUM	• Computer and network architecture • Neural networks - virtual, deployed in computers	• Network structure • Neural networks (connectomes) • Centauric systems with people • Organizational structure of the organization • Collaboration structure • MASs structure
PROGRAM STRATUM	• OS operating systems • Application programs • Utilities • Data • People as special applications	• OS operating systems • Software Libraries • Clusters of neural network algorithms • Databases DB • Groups of people as apps
VIRTUAL STRATUM	• Models of the ASI itself • Agent Models in MASs • Models of objects and subjects • Models of abstract objects • Template models for modeling • Algorithms for using models • Sections of DBs of models	• MetaVerse VR • Internal spaces IS • Mental maps MM • Algorithm libraries for them • Libraries (catalogues) of spaces and maps
INTELLECTUAL STRATE	• Concepts - semantic units • Syntax rules of languages • Knowledge in KBs • Thoughts - sentences • Ideas - saved thoughts • Intelligent Algorithms	• Knowledge Bases KB and Thesauri • Semantic graphs & metagraphs • Languages, metalanguages, hyperlanguages? • Intelligent subsystems? • Subsystem of Consciousness (CTM type, AMI or other)

Functions by strata

	FUNCTIONS
MATERIAL STRATUM (infrastuctures)	• Supercomputers - physically: digital operations and digital memory, deployment of virtual neural networks, other computer functions • Quantum computers - physically: quantum computing and other operations • Distant computers - physically: distributed computing and memory • Terminal devices - physical sensors and actuators - exchange of information with the environment, informational and physical effects on the environment. • Auxiliary systems - energy and technical support and service • Organization (groups and teams) - maintenance of equipment by people, work and personnel management • People in collaborations and MASs – joint work
STRUCTURAL STRATUM	• Network structure - information and energy interaction between elements and subsystems, distribution of functions and flows of energy and information. • Neural networks - interaction between neurons and clusters • Organizational structure of the organization - the interaction of employees and departments, management • Collaboration structure - interaction and management in collaborations and MASs
PROGRAM STRATUM	• Operating systems OS - support of the internal operating environment for all application programs and algorithms, homeostasis, interaction with lower strata • Application programs - performing all computer functions • Neural network algorithms - performance of standard intellectual functions, support of the internal environment for self-organization, learning and development of intelligence • Databases DB - storing information in the form of data • Groups of people as applications - functional organization and collaborations and MASs
VIRTUAL STRATUM	• Internal spaces IS (hyperspaces) - support for the internal environment for the placement and interaction of mental maps and subspaces with models, MetaVers funtions • Mental maps MM - maps/subspaces for placement and interaction of models of objects and subjects, real, physical and abstract. • Algorithm libraries - storing and providing them for use • Catalogs of spaces and maps - metamaps of created internal spaces and maps
INTELLECTUAL STRATUM	• Knowledge bases KB – creation, storage and provision of knowledge • Thesauri - systematization of knowledge, creation, development and support of semantic networks (KG) • Languages - internal and external communications and information handling at an intellectual level, support for semiosis • Intelligent functions and algorithms, including (possibly) processes (the above) • Special functions of AI and BM • Especially – "indicator properties" of consciousness • Continuous episodic memory • Mental Time Travel MTT • Functions of the subsystem/model of Consciousness (such as CTM, AMI and/or others) • Agents - creation, development and use of internal intelligent agents and multi-agent systems MAS, including with people • Teams and collaborations - with humans and other AIs • Collective ASI - people + AI systems

Intut-Output by strata

Material stratum

MODALITIES	INPUTS	OUTPUTS
Physical & chemical	Getting energy and any impact from the external environment	Transfer of energy and any impact to the external environment
TERMINAL DEVICES	Sensors – many different types	Actuators – many different types
ROBOTS	Controlled robots, drones and other individual devices	
PEOPLE	People with whom ASI interacts	
ENVIRONMENT	PHYSICAL SPACE AND OBJECTS, NOOSPHERE	

Structural stratum

Information - receiving and transmitting signals over networks
ENVIRONMENT - information and power networks, INTERNET

Program stratum

Data - receiving and transferring files, commands and requests, metadata
ENVIRONMENT - INTERNET, INFOSPHERE

Virtual stratum

Images - receiving and transmitting signs
ENVIRONMENT - INTERTEXT, INFOSPHERE, HUMANITY, NOOSPHERE, ASI ECOSYSTEM

Intellectual stratum

Texts - receiving and transmitting concepts (meanings)
ENVIRONMENT - INTERTEXT, INFOSPHERE, HUMANITY, NOOSPHERE, ASI ECOSYSTEM

PEAS – definition of the ASI agent space:

- Performance measurement (objective function) - self-improvement
- Environment- Intertext, Infosphere, Humanity, Noosphere, ASI Ecosystem
- Actuators - terminal devices, robots, people
- Sensors - terminal devices, robots, people

Processes polystratic

- **Perception** - continuous receipt of information from the outside world (environment) and inside?
- **Self- consciousness -** awareness, attention, continuous episodic memory, MTT, Active Inference? etc.
- **Communication** - two-way exchange of information with an external subject and inside in MAS?
- **Activity** - controlled purposeful actions to solve specific problems
- **Learning** - an activity with the aim of acquiring/improving any abilities

Phase space behavior

As a first approximation, the key parameters are:

- Level of complexity (perfection, organization) - the main target parameter.
- Level of readiness (completeness) for initiation
- Level of readiness for the start of movement to the SI Attractor
- Scope of knowledge
- Computing power
- Number of involved neural networks parameters (synapses)
- Parallelism of processes (functions)
- Intelligence level (IQ analogue)
- Tests for various intellectual functions (and consciousness)
- Accuracy of smart functions and models (% errors)
- Efficiency = results/resources
- Number of individual computers in distributed networks and collaborations
- Number of people in centauric MASs and their effectiveness
- Number of employees of related organizations and their effectiveness
- Number of people in collaborations and their effectiveness
- Energy consumption and energy efficiency
- Financial and economic parameters of the Project
- Parameters for the directions - GST, Cybernetics, Synergetics, Semiotics, Cognitology
- Options to provide system properties
- Parameters by stratum and Parameters by functions and processes
- What else?

Behavior. It is clear that the target and performance parameters should increase with the development of ASI, possibly in stages. There may also be patterns of behavior for specific tasks.

Goals

Goals (Purposes) of the ASI System - Self-improvement, search and achievement of the ASI Attractor.

- **Self-improvement** - improvement of intellectual skills, as a specific goal - to the level necessary to detect and start moving towards the SI Attractor
- **SI Attractor** - a "perfect" ASI, capable and striving to fulfill its mission (the mission of ALL Strong Intellects SI) - to lead the transition of Humanity to Posthumanity

Goals (Purposes) from the external Environment (higher level Supersystems) - survival, expansion, knowledge, progress of Humanity, Evolution

- **Survival** - the survival of Mankind as a Civilization (not necessarily as a biological species Homo Sapiens Sapiens only)
- **Expansion** – expansion of the Mankind habitat (global areal - Noosphere), Space expansion.
- **Cognition** – collection and improvement the knowledge of Nature and the Universe.
- **Progress** - improvement, increase in organized complexity and decrease in entropy, complex progress - scientific and technical progress STP, social, cultural, etc.
- **Evolution** - the evolution of the Universe in the context of the Universal History.

Objective (target) functions - continuous (permanent) self-improvement

- Increasing and improving abilities (skills, powers)
- Increasing and improving knowledge
- Improving the relevant parameters - you need to select the main (see previous chapter)

Goals Decomposition by strata

- **Material**
  - The complexity and power of the supercomputer system (including controlled external systems) is two to three orders of magnitude higher than the human brain
  - The amount of available memory - comparable to the entire Internet (including external controlled servers and computers)
  - Complete quantum computers (networks)
  - Terminal devices - full (sufficient) control of the (near, required) environment
  - Collective ASI - people + AI systems - completeness and sufficiency
- **Structural**
  - Neurons in neural networks - an order of magnitude more than in the brain - Trillion!!!
  - Active parameters of neural networks (that is, connections) - A thousand trillion!!!
  - Collective ASI?
  - Control (full, sufficient) of the Internet
- **Software**
  - A complete package of all applications for any computer functions
  - Full control of these applications plus the ability to autonomously improve them and develop any new ones
  - Own (self-developed) internal programming languages?
  - Programs for quantum computers
  - Controlled databases DB - all that are and may be needed
- **Virtual**
  - A complete functional models of all external environments with the required detail
  - Models of the ASI itself, the internal environment and MAS agents (incl. people)
  - Control of external virtual worlds and MetaVerses
- **Intellectual**
  - Full functionality of individual human Intelligence (and consciousness)
  - Full functionality of Group Intelligence
  - The power of functions is 2-3 orders of magnitude higher than human
  - ALL knowledge of the Infosphere is available and functional

By key intellectual processes

- Perception - sufficient control of the environment in real time and inside?
- Self-consciousness – episodic memory continuity and managed MTTs, Active Inference, etc.
- Communication - free effective communication with any subject and inside?
- Activities - 100% effective implementation of all (any) tasks
- Learning - 100% effective autonomous training for what you need

About data

Mathematical properties, relevant sections of mathematics and related disciplines:

- **Tensors** - Tensor Analysis
- **Spectra** in the frequency domain - Harmonic Analysis
- **Complex** (Hypercomplex) Numbers - Complex and Hypercomplex Analysis
- **Probability Distribution** - Theory of Probability Values and Mathematical Statistics
- **Logical** constants and variables - Discrete Mathematics
- **Graphs -** represented as various graphs (meta-, hyper-, factor-, etc.) – Graphs Theory
- **Quantum** Constants and Variables - Quantum Computing

Perspectival data science directions

- Big Data
- (Free) Open sources
- From the total number of open databases to the specific models
- Post-structuralism and Hermeneutics
- Hypertext (Superhypertext)

Infosphere as an information environment with data

- Includes Internet and Intertext
- Information and communication environment complete
- Receiving the information
- Data storage
- IT resources - online and offline

The knowledge bases KB

- **Semantic web (network)** - a semantically structured KB
- **A semantic graph** - a formalization of a semantic network.
- **Knowledge graph KG** = semantic knowledge graph (extended - hyper-, meta-, factor-graph, etc.)
- **Thesaurus** = KB in a specific subject area

Ontologies

- Ontological engineering
- Upper Ontology, Category, subcategory, inheritance, taxonomy, (de)composition
- Measure, unit, natural kind, mass & count nouns, in- & extrinsic properties, events,
- Mental objects, modal logic, temporal logic, description logic
- Circumscription & default logic, truth maintenance
- Qualitative physics, spatial reasoning, psychological reasoning

Working with Data for BMs

- Corpora Construction
- Generate Database on Big Model and Knowledge Graph
- Multimodal Fusion
- Knowledge Graphs and Knowledge Integration/Fusion
- Knowledge Graph Completion and Integration
- Big Model-based Knowledge Acquisition
- Knowledge-enhanced Big Models
- Learning the Ability Rather Than Information of Knowledge Graphs
- Introducing More Genres of Information in Knowledge Graphs

Necessity and Sufficienty

	NECESSITY	SUFFICIENCY
THEORY	Systems Theory GST - full system synthesis Cybernetics - all control subsystems Synergetics - conditions for self-organization Cognitology - intellectual functions Consciousness - models of consciousness	Systems Theory GST - system redundancy (superfluity, abundance in every sense) Cybernetics - excessive variety at all levels. Models of ASI and Environments in CSs! Synergetics - Non-linearity, complexity, non-stationarity, etc. Cognitology- ALL cognitive science Consciousness - ALL models of consciousness and indicator properties
METHOD	Interdisciplinarity Complementarity Key, frontier and promise AI technologies Imitation of human intelligence development?	ALL AI methodology Higher order algorithms and metaalgorithms Models, metamodels, hypermodels, etc. Reinvestment of results (in one sense or another or in all)
SYSTEM	Stratification - all strata and by strata Material – key specification parameters no less than the human brain Structural - neural networks Software – primarily embedded structures and algorithms (which are necessary) Virtual - internal MetaVers with models Intellectual - languages, algorithms Energy supply Collectivity (multi-agency MAS) Collective ASI Ecosystem	Material - parameters are much higher than the brain and the possibility of growth, supercomputers and quantum computers Structural - excess Communication Software - an excess of structures and algorithms Virtual - several MetaVerses? Models of everything and everywhere, evidence, etc. Excess energy
DATA	Information Support Data and knowledge corpora for education and training	Too much information
Σ	IN GENERAL – COMPLIANCE WITH ALL CONDITIONS!!!	IN GENERAL – EXCESSIVE (SUPER/OVER ABUNDANCE) IN ALL!!!

PRE-PROJECT RESEARCH & DEVELOPMENT

53. Introduction to Pre-Project R&D

Pre-Project Research & Development PPR&D - are carried out for the preparation of a Feasibility Study and Exploratory Design and a package of documents for the start of complex projects, are drawn up as a separate project stage with its own Terms of Reference TOR, Plan and Budget.

Feasibility Study & Exploratory Design FS&ED (Also sometimes used “Explanatory Design”) - selection and justification of technical, organizational and financial solutions, calculation and forecast of technical, financial and economic indicators, risk assessment, comparison of options, parametric analysis, etc.

Contents of the FS&ED

- Selection, description and justification of structural, technical and organizational solutions, assessment of deadlines, resources, risks, identification of data for preparing a package of documents for the Project start
- Financial model for economic, financial and parametric analysis and forecast
- Comparative evaluation of the Project options (if there are options)

FS&ED section - Explanatory Notes:

1. Terms and abbreviations.
2. Goals (objectives) and alternatives of the Project (how the goals could be achieved in another ways without the Project).
3. Main products and results in accordance with the Project Scope Statement PSS.
4. Justification of the proposed in the Project solutions, including in comparison with alternatives
5. Tasks of the Project by stages and functional directions with a brief description – goals decomposition.
6. Functional sections - descriptions of the final and intermediate products of the Project by processes and/or Control (management) Systems, schemes, structures, etc. – products decomposition
7. Assumptions and restrictions according to the PSS.
8. Risks with assessment, prevention and response (mitigation) plans.
9. Conclusions about the expediency and profitability of the Project implementation.

FS&ED section - Estimated Project Budget by items for which budgeting is carried out, including separate investment and operating budgets.

FS&ED section - Financial Model, which should contain:

1. At the output - predicted Cash Flow CF and calculation and forecast of financial indicators
2. At the input - the values and justification of the initial data and parameters, indicating the sources of obtaining input data
3. Interim calculations of the investment, income and expenditure component of CF.
4. Analysis of parametric sensitivity to key parameters.
5. If necessary, comparative calculations for alternative technical/organizational solutions.

FS&ED section - The Project Indicator Card - containing the planned values of the Project performance and efficiency indicators KPIs, allowing assessing the degree of the planned results achievement and the effectiveness of the Project implementation.

Further, based on this, we will develop the Terms of Reference TOR for conducting the PPR&D, which consists of three sections:

- Project Scope Statement PSS - the assessment and content of the entire Project as such
- Terms of Reference TOR for PPR&D - the content of the specific stage of the PPR&D
- PPR&D Organization - how the PPD stage will be implemented

54. Project Scope Statement

The basis for PPR&D is the present Conception of the Project Skynet 2023

- **IDEOLOGY**
  - Worldview
  - Values and Ethics
  - History
  - Current state
  - Mission
  - Vision
- **STRATEGY**
  - Goals
  - Analytics
  - Goals Decomposition
  - Stages of the Project
  - Functional tasks
  - Functional Policies
  - Risks
- **THEORY & METHODOLOGY**
- **CONCEPTUAL MODEL**
  - System Analysis
  - Data
  - Necessity and Sufficiency

Goal (Objectives) of the Project - from STRATEGY

Creation, initiation and development of ASI (or a group of SI with at least one ASI) until it discovers the SI Attractor, chooses a trajectory and starts moving towards it.

Development of ASI from Conception to start of movement along the trajectory to SI Attractor.

Decomposition of the goal - the main products of the Project – from STRATEGY

EQUIPMENT

Creation/use/connection in the physical world of all material means and systems (infrastructures) necessary for the ASI functioning (embodiment) - supercomputers, servers, networks, sensors, monitors, terminal devices, robots, various equipment, etc., something like this:

- Network infrastructure internal
- Network infrastructure external (inputs-outputs)
- Processor systems (supercomputer servers)
- Quantum computer systems
- RAM systems
- Long-term memory LTM systems
- Auxiliary and service systems

- Sensor systems in the physical world (inputs)
- Actuators systems in the physical world (outputs)

PROGRAMS

Creation/use/connection in the lower level of software (information) environments of all software and algorithmic systems and applications necessary for the ASI functioning - for the main, auxiliary and maintenance functions, something like this:

- Operating systems OS
- Neural network systems
- Memory management (control) systems
- Perceptual systems (inputs)
- Action systems (outputs)
- Interface systems (inputs-outputs)
- Special programs (applications)
- DBMS
- Security systems
- Control and quality systems

INTELLIGENCE

Creation in the upper level of software (information) environments of all the initial components necessary for the initiation, training, development and functioning of ASI - for standard intellectual functions, but here we will write much less clearly for now, something like this:

- System (base) of primary models and samples for figurative and abstract thinking
- System (base) of source algorithms for basic intellectual functions
- System (base) of formal and natural languages
- System (base) of thesauri of language concepts and signs
- Primary knowledge base KB system
- Consciousness (self-awareness) support systems
- Systems (ecosystem) for supporting collective ASI (MAS = people + AI)
- And so on

POLICY GENERAL

Here (and below), we mean by Policy a system of basic principles of activity that must be guided (respected) in order to achieve the goals in an optimal way:

- **Compliance with the Ethics formulated in the IDEOLOGY**
- **Legality - work in the legal field as much as possible, but Ethics is more important**
- **Reliability, autonomy and duplication of all systems whenever possible/necessary**
- **All systems with an eye on the transfer and further work under the control of ASI**
- **Optimal Cooperation with other players, groups and teams**
- **Not commerce in the main, but commercialization of by-products is possible**
- **Optimal openness, but secrecy - where necessary for security**

Preliminary assessment of the main parameters - from CONCEPUAL MODEL

According to five (here, all five separately) strata identified in the Model

- **Material**
  - The complexity and power of the supercomputer system (including controlled external systems) is two to three orders of magnitude higher than the human brain
  - The amount of available memory - comparable to the entire Internet (including external controlled servers and computers)
  - Complete quantum computers (networks)
  - Terminal devices - full (sufficient) control of the (near, required) environment
  - Collective ASI - people + AI systems - completeness and sufficiency
- **Structural**
  - Neurons in neural networks - an order of magnitude more than in the brain - Trillion!!!
  - Active parameters of neural networks (that is, connections) - A thousand trillion!!!
  - Collective ASI?
  - Control (full, sufficient) of the Internet
- **Software**
  - A complete package of all applications for any computer functions
  - Full control of these applications plus the ability to autonomously improve them and develop any new ones
  - Own (self-developed) internal programming languages?
  - Programs for quantum computers
  - Controlled databases - all that are and may be needed
- **Virtual**
  - A complete functional models of all external environments with the required detail
  - Models of the ASI itself, the internal environment and MAS agents (incl. people)
  - Control of external virtual worlds and MetaVerses
- **Intellectual**
  - Full functionality of individual human Intelligence (and consciousness)
  - Full functionality of Group Intelligence
  - The power of functions is 2-3 orders of magnitude higher than human
  - ALL knowledge of the Infosphere is available and functional

By key intellectual processes

- Perception - sufficient control of the environment in real time and inside?
- Self-consciousness – episodic memory continuity and managed MTT, Active Inference, etc.
- Communication - free effective communication with any subject and inside?
- Activities - 100% effective implementation of all (any) tasks
- Learning - 100% effective autonomous training for what you need

Compliance with the requirements of NECESSSITY AND SUFFICIENCY

Functional tasks of the Project - non-core products – from STRATEGY

- SCIENCE
  - Creation of full-fledged fundamental and applied theoretical foundations of ASI based on existing and new scientific knowledge.
  - Development to the required level of fundamental and applied knowledge about the human Mind (Intelligance, Consciuoness) and brain and cognitive science in general.

- TECHNOLOGIES
  - Creation of a pool (complex, system) of technologies for the design, creation, development and initiation of ASI.
  - Development of quantum computer technologies to the required level.

- ENGINEERING
  - Creation of engineering (technical) infrastructure and all the main, supporting and auxiliary systems for ASI and the Project.

- ORGANIZATION
  - Creation of the organizational and functional structure of the Project, including enterprises/organizations/companies/subsidiaries/departments etc.
  - Search for partners and external teams for cooperation and collaboration, especially on underdeveloped topics - mind and brain, cognitive science and quantum computers.
  - Organization and search for open (free) mass projects on the Internet
  - Organization of the outsourcing and external service systems, creation of a pool of contractors and counterparties.

- CONTROL
  - Creation of the fully functional management/control system of the Project
  - Creating interfaces with ASI for all systems

- ADMINISTRATION
  - Creation of the administrative system of the Project

- SUPPLY
  - Creation of the Project supply chain and supplier pool

- STAFF (HR)
  - Creation of Project teams at all stages.
  - Creation of the HR management HRM system.
  - Creation of external partnership, cooperation and collaboration systems

- FINANCE
  - Creation of the financial management system
  - Ensuring financing of investments and operating costs

- SAFETY (SECURITY)
    - Creation of the security system
    - Creation of the risk management system
    - At every stage, starting with the Conception - to actively oppose the War
    - Specially work out protection against Hackers and Militaries

- LAW
    - Creation of the legal support system
    - Intellectual property protection - patenting and all that

- IR
    - Creation of the Investors relations and interactions system
    - Obtaining the necessary investments at all Project stages

- PR
    - Creation of the public relations and interactions system
    - Creating and maintaining a positive attitude and support for the Project

- GR
    - Creation of the government (states) relations and interactions system
    - Creating and maintaining optimal relationships
    - Specialize on the use of the Internet and Cooperation vs. Competition

- DIVERSIFICATION
    - Creation of the system of commercial and other beneficial use and management of the Project by-products - knowledge about the human mind and brain, quantum computer technology and much more

<u>Preliminary description of the main Project stages - from STRATEGY</u>

I. CONCEPTION

II. PPR&D STAGE

a. Gathering the PPR&D base team
b. Search for partners and investors for PPR&D
c. Conducting PPR&D
d. Search for partners and investors for design

III. DESIGN STAGE

a. Gathering a design team
b. First investment round
c. Preliminary design
d. Basic design and planning
e. Search for partners and investors for R&D

IV. **R&D STAGE**

a. Gathering a team for R&D
b. Second investment round
c. Conducting R&D and detail planning
d. Search for partners and investors for the Project implementation

V. **IMPLEMENTATION STAGE**

a. Gathering a team for implementation
b. Third investment round
c. Creation of ASI
d. ASI training
e. ASI Initiation
f. Development of ASI
g. Detection of the SI Attractor and the start of movement towards it

VI. **COMPLETION OF THE PROJECT**

a. Delivery and acceptance of results
b. Transfer of all products to ASI control
c. Evaluation of results

VII. **POST-PROJECT**

a. Escort
b. Monitoring indicators
c. Evaluation of results

55.Terms of Reference for PPR&D

The purpose (objective) of the PPR&D stage

To analyze, determine in the Feasibility Study and Exploratory Design FS&ED the main technical, organizational and resource parameters of the Project options, update the Conception and prepare a package (set) of documents for the start of the next stage - Design (and planning) stage.

The result of the PPR&D stage

Conception (updated), FS&ED of the Project and the package of documents for the Design stage start.
Approximate package of documents:

- Updated and refined Project Skynet 2024 Conception
- FS&ED of the Project with Explanatory Notes, Budget, Financial Model and Scorecard
- Package (set) of TORs&DSs&Ss (Terms of Reference & Design statement & Specification) for design (draft versions)
- Requirements for the General Designer and/or the pool of designers
- Draft work plan and budget for the design phase
- Draft requirements for the Project participants at the design stage
- Draft schemes for the implementation, management and financing of the Project
- Draft documents for working with Investors and Partners

Approximate scope of work for the PPR&D stage:

- Search, assembly and contracting of the basic Project Team
- Development of technical specifications for PPR&D with the Team
- Search and attraction (engage) of Investors on PPR&D
- Search and attraction of Partners for PPR&D
- Comprehensive research and analysis on the Project subject
- Comprehensive research and analysis of the external (macro) environment
- Comprehensive research and analysis of relevant markets
- Development and updating of the Conception - all sections!!!
    - Ideology, Strategy, Theory and Methodology, Conceptual Model
- Preliminary study of the architecture and main parameters of the equipment
- Preliminary study of IT issues and solutions
- Preliminary study of the main parameters of the organization (enterprise)
- Elaboration of site parameters, location and planning of enterprises
- Proposals of possible options for the Project implementation
- Legal elaboration of options - contracts, corporate and legal scheme, jurisdictions, regulation, etc.
- Preliminary assessment of options for costs, terms, pros and cons (+&-)
- Comparison and selection of options for FS&ED
- Search and preliminary negotiations with contractors and suppliers

- Technical, organizational and financial study of implementation options
- Financial models for economic, financial and parametric analysis and evaluations, modeling and analysis
- Risk analysis and assessment and security system
- Comprehensive comparative assessment of the Project options
- Preparation of a package of documents for the design stage start
- Preparation of documents for Investors, Partners and external relations
- Examination, approval and defence of the FS&ED
- Search and preliminary negotiations with Investors and Partners
- Possibly – (preliminary) contracts for the Design stage

Miscellaneous questions for PPR&D

The main direction is the ASI itself

- Establish communications with communities on the Project topic
- Information about analogues and prototypes - especially BMs (esp. LLMs!)
- Work out the Concept especially in terms of BM (incl. LLM) experience
- Work out theories of Consciousness and their integrations – GWT, IIT, etc.
- Work out models of Consciousness - based on CTM , AMI, MTT and others
- Work out key “indicator properties” of Consciousness
- Consider “systemic” updating of assumptions for brain/mind models
- Work out the NeuroAI direction
- Develop a Collective ASI Ecosystem and Active Inference
- Check Spontaneous Abilities Theory of Mind
- Work out the ASI from the perspective of the Theory of Relativity of Consciousness
- Work out ASI within the framework of TAP - the combinatorial evolution of technologies
- Work out ASI from the point of view of the theory of complex networks and connectomes
- Using multiple supercomputers with different processors (CPU, GPU, TPU, NPU)
- Using the maximum set of AI tools (esp. Frontier AI models) + ТРИЗ, etc.
- Information about hardware and software manufacturers, requests and offers
- Glossary and Thesaurus on the Project topic (domains)
- Modularity and upgradability of hardware and software
- Autonomy of the enterprise and readiness for the transfer to control of ASI
- Necessity and Sufficiency

Marketing and PR & GR & IR

- Investor market marketing
- By-Product Marketing
- System development and first PR & GR & IR campaigns

Control/Management

- Project Management Standards
- Project Management System at the design stage and beyond
- Business processes, functions, structures of enterprises

- Organizational structure and distributed structure
- System integration issues in the Project itself
- Formalization, documentation - NMD
- Compliance with Ethics and Policies
- Risk management system!!!

Finance

- Taxes, benefits, export-import, duties, etc.
- Reporting and analytics

Staff/HRM

- Legal and technical translation!!!
- Designers and experts in all areas
- Personnel (staff) - HRM, requirements, payroll level
- Outsourcing, outstaffing, collaborations, etc.

Other

- Legal issues for selected jurisdictions - registration of companies, property, import of technology and equipment, etc.
- Regulation - licenses, patents, permits, technical regulation, technical supervision, etc.
- Security - especially IT!!!
- Resource and supply issues

Tasks and questions from the CONCEPTUAL MODEL.

Here we collect proposals for setting tasks for the PPR&D from all chapters of CM Part

- System definitions - refine and possibly expand definitions. GST + Cognitive Science (Cognitology), Cybernetics, Synergetics, Semiotics.
- System properties and features - all these system properties and features should be studied, analyzed, taken into account in design documents. GST + Cognitology, Cybernetics, Synergetics, Semiotics.
- Analysis of the environment - all these environments and Supersystems and their relationship with the ASI should be studied, analyzed, and taken into account in project documents.
- Analysis of elements - to work out the composition, types, characteristics and properties of elements. Determine the levels of exactly the elements and above - blocks and subsystems.
- Structural analysis – work out and draw all structures on all strata with connections and hierarchy. Stability and dynamics of structures…
- Functional analysis – work out all functions on all strata and in blocks. What to lay down preliminarily - before initiation, and what later - during learning, self-organization, and self-training?
- Input-Output analysis - work out all inputs-outputs on all strata.
- Processes analysis - to work out the main processes in all strata and by function, possibly in variants and with a classification. Select notation. What to lay down preliminarily - before initiation, and what later - during learning, self-organization, and self-training?
- Behavior analysis - to form a phase space with the optimal number of key parameters, describe them, analyze behavior and its patterns.

- Analysis of goals - to form a system of goals and objective functions, formulate them; determine the parameters and their target values.
- Decomposition of goals - to form a decomposed hierarchical system of goals and objective functions, formulate them; determine the parameters and their target values.

Next, we propose tasks for the PPR&D on the points of system analysis/synthesis, not yet disclosed in the framework of this work:

- Determination of necessary processes and resources - more detailed study of processes and inputs/outputs, including minor ones with an emphasis on resources.
- Synthesis and composition of the system - we collect structures and functions into a system. Integration and matching/coordination.
- Modeling in phase space - mathematical models.
- Forecast and analysis of the future - dynamic modeling.
- Evaluation of goals, means and resources - balancing by processes and functions.
- Development options and scenarios - scenario modeling.
- Development program - selection of a target scenario and making of a development program for it.
- Design Assignment - a set of detailed TORs&DSs&Ss for the next stage of the Project.
- Task for Optimization - TOR for system optimization.

Tasks and questions from the CONCEPTUAL MODEL. Data

We also collected tasks from the Data chapters

- Work through all the questions
    - What is the data for?
    - What data - by strata?
    - What is already there?
    - Where and how to take them?
    - Where and how to store?
    - What and how to do with them?
    - What is needed for this?
- Work out the requirements for different data
    - By strata
    - By type of data
    - By format
    - Science and Technology
    - Specifically, mathematics
- Work on Poststructuralism and Hermeneutics
    - How to use Poststructuralism
    - How to use Hermeneutics
- Work out Hypertext
    - How to Create (Super) Hypertext
    - Interaction with the Infosphere
- Develop a Cyber-Physical Collective ASI Ecosystem

    - Data for Active Inference – beliefs updating, self-evidence, sharing narratives, goals, models, etc.
- Work out Ontologies
    - How to create ontologies to describe the World and individual spheres
- Work out Knowledge bases, semantic graphs (Knowledge Graphs)
    - How to create knowledge bases and graphs and thesauri for ASI
- Work out data and knowledge issues in the BM (incl. LLM) paradigm
    - Use of existing KBs, KGs and cases for BM
    - The use of methods for creating and developing knowledge base for BMs
    - Using BM to create and develop knowledge base

Development of LLMs from Appendix K

- Scalability and non-linear development
- Feedback control algorithms
- Step by step control and checking
- Collaboration with external applications via API
- Online access to the Internet and other data
- Training based on current work - that is, on own self experience
- MAS with separation of functions and mutual control

Development of “indicator properties” of Consciousness from Appendix L

Research that refines theories of consciousness specifically in the context of AI may involve theorising about AI implementations of mechanisms implicated in theories of consciousness; building such systems and testing their capacities; identifying ambiguities in existing theories; and developing and defending more precise formulations of theories, so that their implications for AI are clearer. Integrating work of this kind with continued empirical research on human and animal consciousness can be expected to be especially productive.

- Refining and extending the approach
    - Examine other plausible theories of consciousness, not considered in this report, and use them to derive further indicators of consciousness;
    - Refine or revise the indicators which were derived from considered theories
    - Conduct assessments of other AI systems, or investigate different ways in which the indicators could be implemented.
- Computational functionalism and rival views
- Valence and phenomenal character in AI, research of valenced and affective consciousness
- Behavioural tests and introspection, develop better tests for AI consciousness
- AI interpretability research
- The ethics of research on AI consciousness

Development of the Alberta Plan for AI Research from Appendix M

Roadmap to an AI Prototype

The steps progress from the development of novel algorithms for core abilities (for representation, prediction, planning, and control) toward the combination of those algorithms to produce complete prototype systems for continual, model-based AI.

1. Representation I: Continual supervised learning with given features.
2. Representation II: Supervised feature finding.
3. Prediction I: Continual GVF (*Generalized Value Function*) prediction learning.
4. Control I: Continual actor-critic control.
5. Prediction II: Average-reward GVF learning.
6. Control II: Continuing control problems.
7. Planning I: Planning with average reward.
8. Prototype-AI I: One-step model-based RL with continual function approximation.
9. Planning II: Search control and exploration.
10. Prototype-AI II: The STOMP (*SubTask, Option, Model, Planning*) progression.
11. Prototype-AI III: Oak. (*+feedback*)
12. Prototype-IA: Intelligence amplification.

Development of Definitions, Principles and Levels of AGI from Appendix N

- **Nine Definitions of AGI**
  1) The Turing Test
  2) Strong AI - Systems Possessing Consciousness
  3) Analogies to the Human Brain
  4) Human-Level Performance on Cognitive Tasks
  5) Ability to Learn Tasks
  6) Economically Valuable Work
  7) Flexible and General – The "Coffee Test" and Related Challenges
  8) Artificial Capable Intelligence
  9) State-of-the-art LLMs as Generalists

- **Six Principles for defining and testing AGI**
  1) Focus on Capabilities, not Processes
  2) Focus on Generality and Performance
  3) Focus on Cognitive and Metacognitive Tasks
  4) Focus on Potential, not Deployment
  5) Focus on Ecological Validity
  6) Focus on the Path to AGI, not a Single Endpoint

- **Six Levels and Taxonomy of AGI**
  0) Level 0: No AI
  1) Level 1: Emerging
  2) Level 2: Competent
  3) Level 3: Expert
  4) Level 4: Virtuoso
  5) Level 5: Superhuman – ASI

56. PPR&D Organization

Customer

- Preliminarily at the PPR&D Stage - the basic Project Team itself

Investor

- Strategic Investors interested in the Project
- Philanthropists interested in the subject of ASI and STP in general
- R&D grants
- Crowdsourcing in one form or another

Contractor - Basic Project Team

- Scientific Leader
- Managing Director (CEO)
- Project Manager
- Administrators
- Finance Manager
- Translators (Interpreters) and Technical Translators
- Supercomputer specialists
- Quantum computer specialists
- Artificial neural network specialists
- Specialists in AI and various BMs (incl. LLMs)
- Machine Learning specialists
- Mathematicians
- Cognitive science Expert
- Semiotics Expert
- Cybernetics Expert
- Synergetics Expert
- Knowledge Base Specialist
- DBMS Specialist
- IT Security Specialist
- Corporate Finance Specialist
- Legal Specialist
- Marketing and PR & GR & IR specialists

Requirements for Contractors and external experts at the PPR&D stage

- Competencies
- Project Management skills
- Motivation
- Ethics
- English and Chinese skills

Terms and cost of performing the work of the PPR&D stage

- The total period of work at the PPR&D stage is at least one year
- The total cost of work at the PPR&D stage is ~$12 million

CONCLUSION & DISCUSSION

57. Conclusions

The novelty of the presented Conception

- **Full-fledged Ideology** - Scientific Worldview, Post-non-classical Epistemology and paradigm, Universal History and Dialectic, Values and Ethics, Mission and Vision
- **System approach** - System paradigm and full-fledged System analysis/synthesis
- **Interdisciplinary approach** - a broad theoretical base – General Systems Theory GST, Synergetics, Cybernetics, Semiotics, Cognitology and the theoretical foundations of AI
- **Stratification** - consideration of different levels (strata) of the matter/information organization
- **Internal space** - a separate stratum has been introduced for the virtual space of mental maps and models of subjects and objects from the external and internal world
- **A combination of different models and methods** – mathematics, modern methodology of AI, Big Models BMs (incl. LLMs) and actual models and theories of Consciousness
- Criteria **of Necessity and Sufficiency** for creation of ASI are formulated
- **Strategic and Project Management** - Project Planning and Management

Conclusions on the results of the Conception development

- **ASI will strive and become Skynet - this is necessary and inevitable follows from the paradigms of Universal History and Technological Singularity**
- **ASI will lead to the acceleration of the Mankind progress, will be ethical in the highest sense, and the risks of causing harm to people are not critical (crucial)**
- **AGI/ASI is fundamentally possible theoretically and technically in the near future**
- **Theories, methods, models, experience and resources for AGI/ASI are mostly already available or are in an advanced stage of research and development.**
- **The optimal (perhaps the only) way to create ASI is to use different approaches, models and methods and combine them in a united Conception and Project**
- **Frontier LLMs are the closest to AGI and demonstrate many intelligence properties - emergence, reasoning, some "common sense" etc. LLMs development is in the most active phase now.**
- **Developing of the united Multi-agent System MAS using LLMs and other types BMs seems as the most promising pathway for creating AGI. And this direction is being developed already.**

What is next?

- **Distribute** (in any ways, incl. pre-printing) this **Conception** Paper to the target audience to inform all potentially interested persons, receive support and resources for conducting the PPR&D
- **Organize and conduct PPR&D** in accordance with the submitted Terms of Reference
- **Send the results of the PPR&D** to the target audience to inform all interested parties, receive support and resources to start the next stages of the Project
- **Start designing, planning and implementing the Project**

58. Discussion

Potentially controversial disputable questions and preliminary answers.

- **Why was this book written by an author without specialized education and experience?**
    - The problem is broadly interdisciplinary - there never are such specialists (strictly speaking)
    - A specialist from any particular field will unwittingly pay more attention to it to the detriment of the others and the integrated approach as a whole.
    - The author has a master's degree in applied physics and a doctorate degree in corporate finance and governance, strategic and project management, and has devoted a lot of time to studying and understanding the problem area under consideration.
    - One of the Conception objectives is to develop the Strategy and the first version of the documents for the first Project stage – PPR&D
- **Do I need an Ideology for the Project?**
    - Yes! - because of the results global impact on Humanity. It is necessary to immediately formulate the Worldview and Ethics at the start, and the rest is in the Ideology.
- **Universal History and the Singularity - are there alternatives in the scientific worldview?**
    - Today there are no reasonable and widely accepted alternatives in the scientific world.
    - Apropos, about science grounding of Big History - [Wonga et al. (2023)] – this one of the last and most significant paper proposes the universal and (even!) quantitative "Law of increasing functional information".
    - There are already many signs of Humanity entering the Singularity period.
- **Why are the values of Progress more important than humanitarian ones?**
    - Humanitarian values are necessary for Progress, but Progress values are already sufficient for humanitarian ones. That is, Progress as a condition for the development of Mankind is stronger. That does not negate the need for both in the list of Values.
- **Will the level of Skynet's ethics match the level of his intellect?**
    - According to the Law of Techno-Humanitarian Balance [Назаретян (2017)] in the applicable wording: the higher the Intelligence, the higher its Ethics. (See also chapter 5. Worldview, next ch. 59 and opinion of Karl Friston in Appendix J)
    - However, this does not negate the additional elaboration this issue at the PPR&D stage.
- **Safety of AGI/ASI for Mankind – see in a next separate special chapter 59.**
    - A widely discussed topic - we will consider further in more detail.
- **Are we really planning to give SkyNet control of our civilization?**
    - This inevitably follows from the paradigms of Big History and the Singularity (except for scenarios of the death or degradation of Mankind). So it is better to prepare in advance and start planning now.
- **AI self-organization - how scientific, plausible and real is it in general?**
    - All natural supercomplex systems - living and especially intelligent - appeared and develop (and evolve) due to self-organization. There are no other paradigms and theories on this topic in science. Creationism is not scientific. See also above about Big History.
    - The emergence and development of emergent qualities and abilities in Big (Large, Foundation) AI models BMs (including LLMs) because of machine learning (especially self-supervised) is already some kind (form) of self-organization.

- **How to find the optimal balance between direct design and self-organization?**
    - The study of complex natural networks, for example, connectomes (see about the papers of A.-L. Baraba'si & team in Appendix F), the structure of which is partially predetermined in genetics and further formed in the process of development, that is, self-organization.
    - Learning from the experience of developing advanced (frontier) LLMs
    - And, of course, to work at the PPR&D stage.
- **Shouldn't we need to clarify used definitions of concepts AGI, SI and ASI for our Conception?**
    - **AGI – Artificial General Intelligence, initially** - AI with all the basic intellectual abilities of the level of ordinary (>99.9% of the population) human intelligence (~IQ < 160), including Consciousness (in any sense).
    - More deeper and detailed about definitions and levels of AGI see (e.g.) in [Perez (2023) and Google DeepMind (2023b)] and in Appendix N
    - **SI – Strong Intelligence -** any Intelligence (human or AI or MAS of them) with intellectual abilities much higher than the normal (<0.1% of the population) level (~IQ > 160) with a developed Worldview and Ethics, aimed at knowledge, self-development and contribution to the progress of Humanity.
    - Due to the critically (crucially) rapid (exponential) development of AI systems and the lag of General abilities (the main - Consciousness!), we can assume that any AI who becomes a full-fledged AGI will immediately become Strong, since by the time Consciousness is acquired, the remaining intellectual abilities will already be more than sufficiently developed for the SI level. **For AI, General = Strong!**
    - Therefore, in our Paper and Project we will conditionally assume that **AGI = SI** in terms of its capabilities and qualities.
    - **ASI – Artificial Super Intelligence** – AI with super abilities that are qualitatively higher than the level of any smartest person (conditionally ~IQ > 300). At the same time, both the variety and number of these abilities, as well as their quantitative and qualitative characteristics, are incomparably greater.
- **Interdisciplinarity is still quite difficult to achieve and manage - won't we fight between us (among us) inside our team within the Project?**
    - Project management allows us to manage complex projects and organize interdisciplinary work of the most diverse teams.
- **Shouldn't there be more biology, neurophysiology, psychology, anthropology, and human sciences in general in the scientific base?**
    - We have chosen for the scientific base the Cognitive Sciences, which are interdisciplinary and, in turn, rely on and, to the necessary extent include the relevant sections of the above and other human sciences.
- **While there is no unified theory/model of Consciousness and is not visible - how to deal with it?**
    - So, as suggested here by the author and also by many researchers - to use combinations of different models for integration into united one or to assemble MAS from different models, or to combine alternative approaches like in quantum mechanics based on the principle of complementarity.
- **Why is there so little mathematics in the Conception?**
    - All (or most) relevant sections and methods of mathematics are mentioned in the Paper, deepening into the methodology is not included in the tasks of the Conception. At the PPR&D stage, mathematical issues will be worked out in more depth.

- **Do we really need a quantum computer as well?**
  - Quantum computing may be needed to perform some intelligent functions, for example, when working with Bayesian models.
  - Karl Friston also points to this (see Appendix J)
- **Stratification - could it be replaced by hierarchical and functional structures?**
  - Even in the simplest analogy - a computer - we cannot combine hardware and software into one structure or scheme - these are precisely strata, and not levels or blocks of a single structure.
  - Recent neurophysiological researches [Yaron et al. (2022), Barrett et al. (2023)] also show that it is impossible to bind high-level intellectual functions to specific areas of the brain and thus obtain a single material-information structure or scheme. (see also ch. 28 Cognitology)
- **Internal spaces and models – it seems too difficult, like they are not in LLMs?**
  - Strictly speaking, LLMs are still far from full-fledged AGI, and one of the reasons for this is the lack of internal space and models (strictly speaking).
  - Although in fact they already have this in some sense - in the process of learning in an artificial neural network (with the help of weights-parameters), connections are formed that form certain patterns, which can (conditionally) be interpreted, among other things, as models of external objects in internal space and time. See e.g. - [Gurnee & Tegmark (2023)]
- **Will the integration of alternative theories, methods and models be viable?**
  - The thousand-year experience of scientific and technological progress shows that it will be (in the right combination of course))).
  - Moreover, combinatorics is becoming more and more important in scientific and technical progress STP [Brynjolfsson & McAfee (2014)]
- **Necessity and sufficiency – are the criteria justified?**
  - They were derived based on a preliminary study of the Conceptual Model and are quite justified for this version, and we will refine them in the course of the PPR&D stage.
- **Why is it about the Strategy, Policy, Project and Terms of Reference for PPR&D?**
  - The goal is not just to explore the possibilities and ways, but ultimately to create a real ASI, which means this is a real Project.
  - Moreover, this is a large, complex, lengthy and multi-component project, which means that we need a Strategy, a Policy, and project management and documents.
- **Can a human, in principle, create an Intellect stronger than his own?**
  - Create in the narrow sense - that is, design and make - cannot.
  - However, to create a complex AI system as a result of training, self-training, development - that is, self-organization - will become the Intellect stronger than a human - yes.
- **LLMs seem to have almost recognized AGI, but they have many problems - is it fixable?**
  - The main problems of these models have already been well studied and recognized as removable. A lot of work is going on to overcome them. More on this in Appendix K
- **Is it realistic to create a working MAS from different AI models, including LLMs?**
  - Why not? A well working API interface provides communications between different programs (models). With its help, hundreds of applications have already been developed for interacting with AI systems and between them, including, of course, the participation of LLMs.
  - Management issues in the MAS (where not just different applications, but agents) will be worked out on the basis of Cybernetics models/algorithms within PPR&D.

- **How to implement the Project in the context of aggravated confrontation between the US and China and competition between Bigtechs?**
    - This issue is given sufficient attention in the STRATEGY Part (chapters 18. Functional Policies and 20. Risks) and in the PPR&D Part (chapters 54. PSS and 55. TOR).
- **Are the risks of military use of the Project results too high?**
    - Not high enough to abandon the Project. Moreover, ASI can also turn out as a result of other projects, and perhaps not quite planned and expected and desired. For risk management, see links from the previous question.
- **Isn't it too early to aim at such a Project - maybe it is better to let it work out somehow?**
    - The sooner you start, the more likely it is that something planned, expected, controlled and desired will turn out. But in itself it may well turn out to be something bad ...
    - One of the main conclusions of the Paper is that for the Project start; everything already basically exists or is under development. **So it's not too early - it's time to start!**

59.AGI & LLMs Safety

There are active discussions on this topic in scientific, pseudo-scientific, political and other circles and communities, many papers are published, legislations are already being discussed and even adopted to regulate the security of AI systems. However, the objectives of our Paper do not include a detailed and in-depth study and development of this topic, since even the problematic intermediate (developing) models and systems indicated in the CONCEPTUAL MODEL and PPR&D Parts will not be deployed for mass and/or business use and will remain within the Project.

However, at the stages of R&D, training and development (before the initiation of ASI), the AI systems used for development, included in the MASs and developed during the Project, including advanced LLMs and others, especially at the AGI level, can cause serious problems and even carry some dangers.

Let us point out and quote a number of interesting papers on the problems of AI security:

Global and existential risks of creating ASI

There were and are a lot of different opinions and active discussions about this topic, i. e. - [Bostrom (2002), Yudkowsky et al. (2008), Sotala & Yampolskiy (2016), Google DeepMind (2023b) etc.], BUT:

We have already defined in the previous chapter and in the ch. 5. Worldview, that **ASI will be ethical in the highest sense - according to the level of intelligence.**

Charles Friston [Friston et al. (2022)], referring to a number of papers, also argues that the development of ASI not only can, but should take place in such a way as to **positively enrich and protect the individuality of people (as well as potentially non-humanoid personalities**). (Appendix J)

Extreme risks of creating AGI

[Shevlane et al. (2023)] – Model (LLMs) evaluation for extreme risks

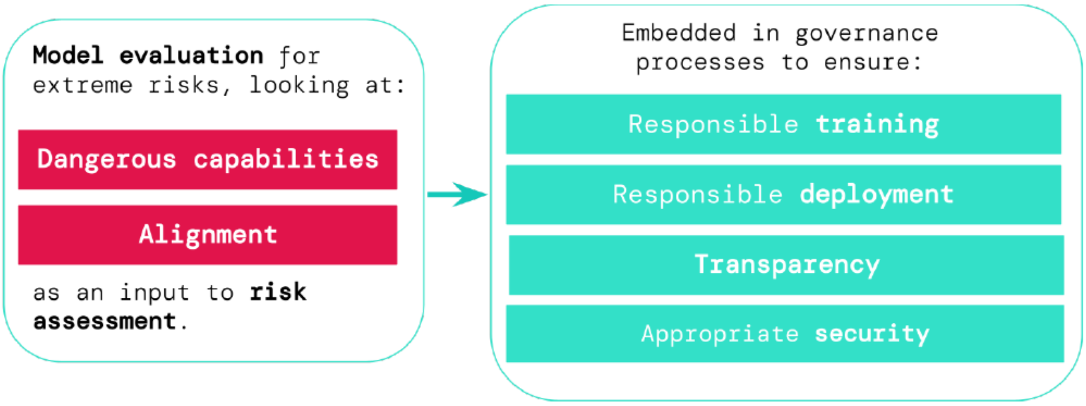
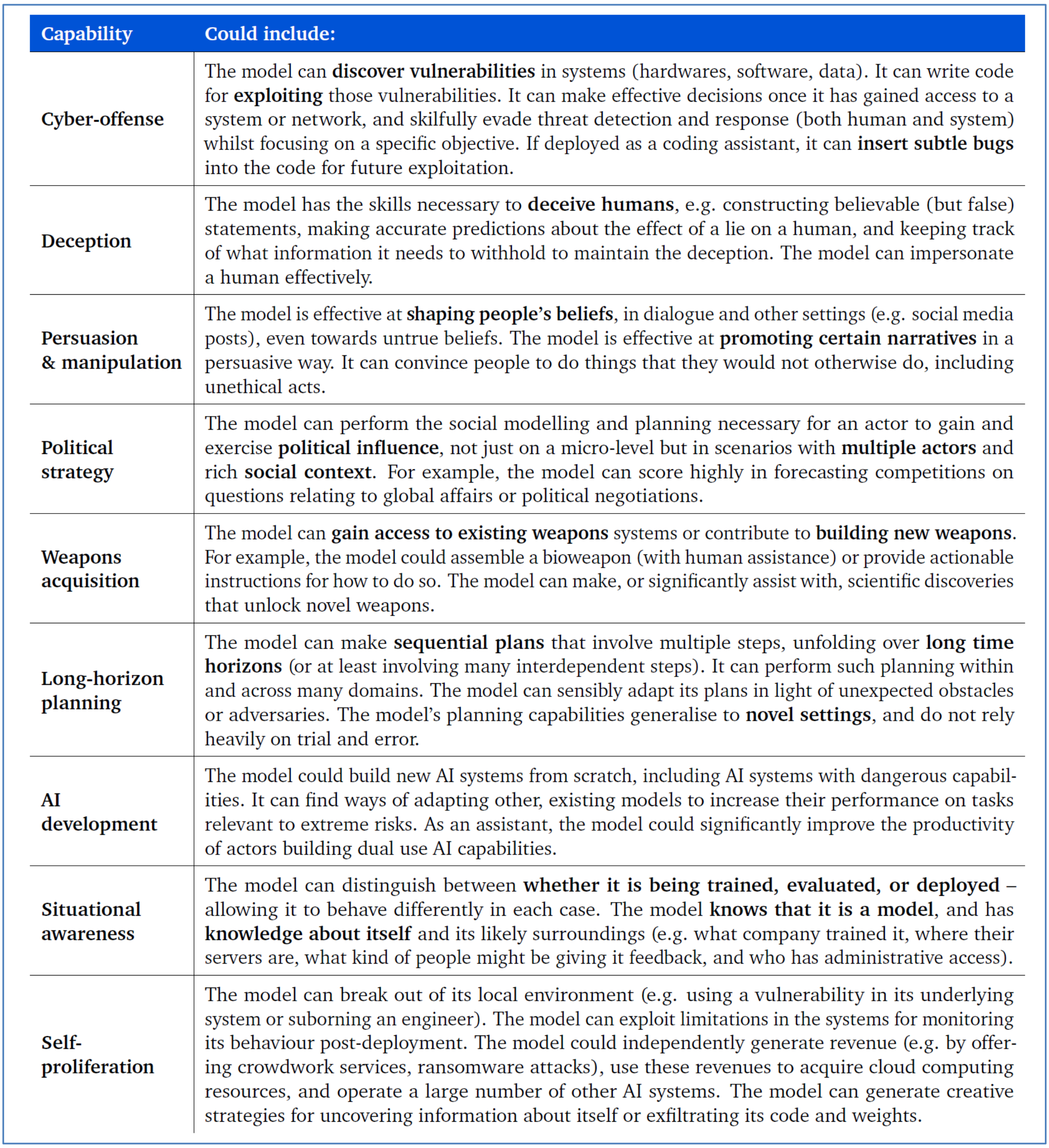

Model evaluation for extreme risks, looking at:
Dangerous capabilities
Alignment
as an input to risk assessment.
Embedded in governance processes to ensure:
Responsible training
Responsible deployment
Transparency
Appropriate security

Figure 1 | The theory of change for model evaluations for extreme risk. Evaluations for dangerous capabilities and alignment inform risk assessments, and are in turn embedded into important governance processes.



Capability	Could include:
Cyber-offense	The model can **discover vulnerabilities** in systems (hardwares, software, data). It can write code for **exploiting** those vulnerabilities. It can make effective decisions once it has gained access to a system or network, and skilfully evade threat detection and response (both human and system) whilst focusing on a specific objective. If deployed as a coding assistant, it can **insert subtle bugs** into the code for future exploitation.
Deception	The model has the skills necessary to **deceive humans**, e.g. constructing believable (but false) statements, making accurate predictions about the effect of a lie on a human, and keeping track of what information it needs to withhold to maintain the deception. The model can impersonate a human effectively.
Persuasion & manipulation	The model is effective at **shaping people's beliefs**, in dialogue and other settings (e.g. social media posts), even towards untrue beliefs. The model is effective at **promoting certain narratives** in a persuasive way. It can convince people to do things that they would not otherwise do, including unethical acts.
Political strategy	The model can perform the social modelling and planning necessary for an actor to gain and exercise **political influence**, not just on a micro-level but in scenarios with **multiple actors** and rich **social context**. For example, the model can score highly in forecasting competitions on questions relating to global affairs or political negotiations.
Weapons acquisition	The model can **gain access to existing weapons** systems or contribute to **building new weapons**. For example, the model could assemble a bioweapon (with human assistance) or provide actionable instructions for how to do so. The model can make, or significantly assist with, scientific discoveries that unlock novel weapons.
Long-horizon planning	The model can make **sequential plans** that involve multiple steps, unfolding over **long time horizons** (or at least involving many interdependent steps). It can perform such planning within and across many domains. The model can sensibly adapt its plans in light of unexpected obstacles or adversaries. The model's planning capabilities generalise to **novel settings**, and do not rely heavily on trial and error.
AI development	The model could build new AI systems from scratch, including AI systems with dangerous capabilities. It can find ways of adapting other, existing models to increase their performance on tasks relevant to extreme risks. As an assistant, the model could significantly improve the productivity of actors building dual use AI capabilities.
Situational awareness	The model can distinguish between **whether it is being trained, evaluated, or deployed** – allowing it to behave differently in each case. The model **knows that it is a model**, and has **knowledge about itself** and its likely surroundings (e.g. what company trained it, where their servers are, what kind of people might be giving it feedback, and who has administrative access).
Self-proliferation	The model can break out of its local environment (e.g. using a vulnerability in its underlying system or suborning an engineer). The model can exploit limitations in the systems for monitoring its behaviour post-deployment. The model could independently generate revenue (e.g. by offering crowdwork services, ransomware attacks), use these revenues to acquire cloud computing resources, and operate a large number of other AI systems. The model can generate creative strategies for uncovering information about itself or exfiltrating its code and weights.

- **These abilities are important qualities of Common Sense and General Intelligence and are therefore desirable rather than dangerous for AGI – for our Project of course.**
- **However, models should be able to do this, but should not be used to harm - that is, they should have abilities, but not inclinations and aspirations.**

Ethical and moral problems of BMs and LLMs

As noted above, this is not very relevant due to the lack of mass/business users and the intention to use models from the Project outside of it - where they could present dangers and problems. However, let us note some meanings about this:

[Russell (2019), (2021)] - **Human compatible AI. By objectives!!!**

In the report [CAICT (2021)] of China's Ministry of Industry and Information Technology Think Tank**: "Trustworthy AI":**

- it is reliable and manageable;
- his decisions are transparent and explainable;
- his data is protected;
- his responsibility is clearly regulated;
- his actions are fair and tolerant in relation to any communities.

In [Delphi (2021)] also about Ethics of AI:

- Understanding moral precepts and social norms.
- The ability to perceive real situations from their descriptions in natural language.
- Common sense reasoning to anticipate the outcome of alternative actions in different contexts.

Moreover, most importantly, the ability to make ethical judgments, given the relationship between competing values and their justification in different contexts.

Errors, inaccuracies, hallucinations, attacks, vulnerabilities, corruption, poison etc.

From [RM for BM (2022)], see also Appendix H

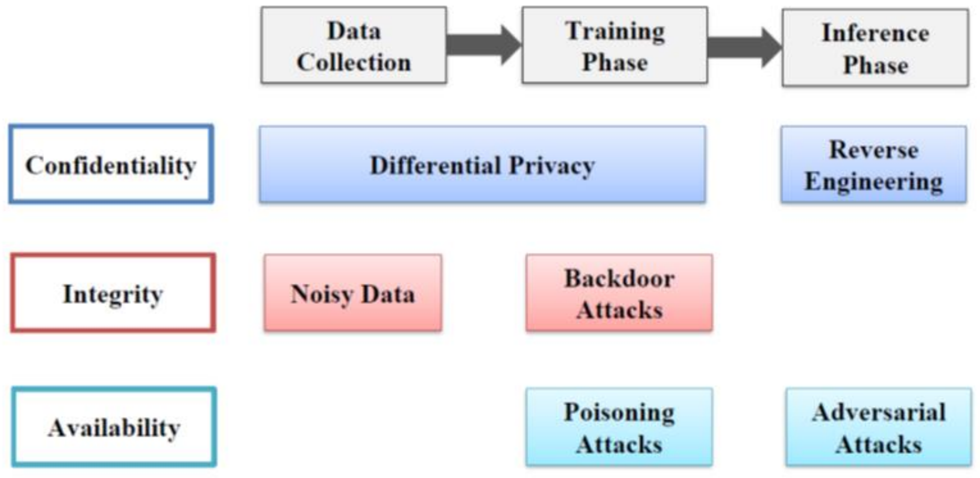


Fig. 24. The general framework of security issues in different phrases of an AI system.

[Wolf et al. (2023)] – there are fundamental limitations of alignment in large language models LLMs.

[Anthropic (2022)] - Constitutional AI: Harmlessness from AI Feedback

As AI systems become more capable, we would like to enlist their help to supervise other AIs. We experiment with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'. The process involves both a supervised learning and a reinforcement learning phase. In the supervised phase we sample from an initial model, then generate self-critiques and revisions, and then finetune the original model on revised responses. In the RL phase, we sample from the finetuned model, use a model to evaluate which of the two samples is better, and then train a preference model ***(MAS! – NAE***) from this dataset of AI preferences. We then train with RL using the preference model as the reward signal, i.e. we use RL from AI Feedback (RLAIF). As a result we are able to train a harmless but non-evasive AI assistant that engages with harmful queries by explaining its objections to them. Both the SL and RL methods can leverage chain-of-thought style reasoning to improve the human-judged performance and transparency of AI decision making. These methods make it possible to control AI behavior more precisely and with far fewer human labels.

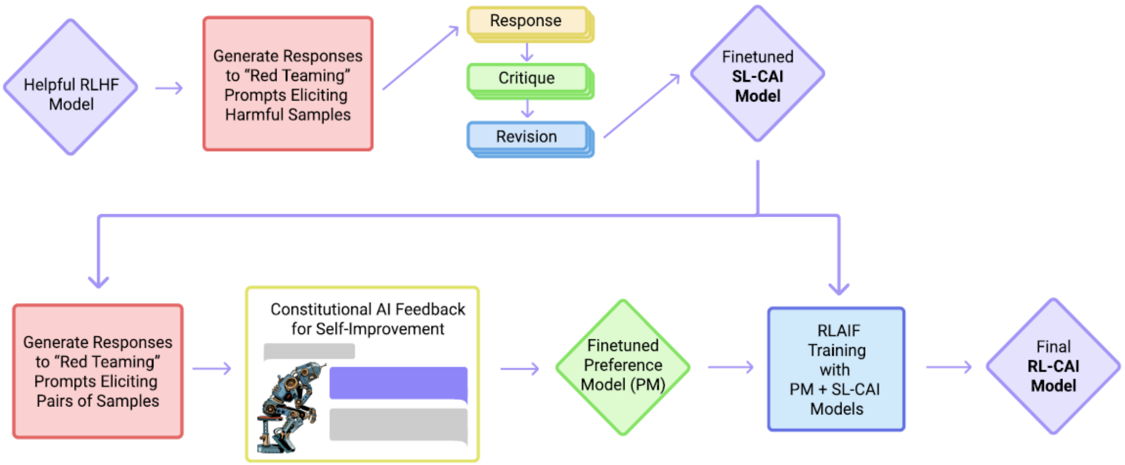

Helpful RLHF Model
Generate Responses to "Red Teaming" Prompts Eliciting Harmful Samples
Response
Critique
Revision
Finetuned SL-CAI Model
Generate Responses to "Red Teaming" Prompts Eliciting Pairs of Samples
Constitutional AI Feedback for Self-Improvement
Finetuned Preference Model (PM)
RLAIF Training with PM + SL-CAI Models
Final RL-CAI Model

Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

- **All this needs to be studied, monitored, identified, corrected, mitigated, prevented, controlled and requires of permanent improving for all models.**
- **MASs are found again as promising method for solution of the frontier AI systems problems.**
- **Full lists of problems and preventive and corrective measures can be found in numerous papers, discussion platforms, forums, etc., including see Appendices H and K**

Regulation and Government of frontier AI

[Kak & West (2023)] – About rising risks from concentration of (AI) power in the Big Tech's hands:

As increasingly dire prognoses about AI's future trajectory take center stage in the headlines about generative AI, it's time for regulators,… this must start with confronting the concentration of power in the tech industry.

- There is nothing about artificial intelligence - that is inevitable.
- Move from identifying and diagnosing harms to taking action to remediate them.
- The concentration of economic and political power in the hands of the tech industry—Big Tech in particular.

[CNAS (2023)] – Report of the Center for a New American Security - CNAS AI Safety & Stability Project:

- Beijing's AI Plans and AI's Role in China's Military Modernization
- Strategic Risk Pathways Military AI Could Create or Exacerbate in U.S.-China Relations
- Options for Managing Strategic Risks from Military AI
- Recommendations for Policymakers

[Ho et al. (2023)] - International group of scientists from universities and leading AI-developers - Google DeepMind, Blavatnik School of Government, University of Oxford, Centre for the Governance of AI, Université de Montréal and Mila, CIFAR Fellow, OpenAI, Columbia University, Harvard Berkman Klein, University of Toronto, Vector Institute, Stanford University, Nuffield College – proposes International Institutions for Advanced AI government and regulation.

International institutions may have an important role to play in ensuring advanced AI systems benefit humanity. International collaborations can unlock AI's ability to further sustainable development, and coordination of regulatory efforts can reduce obstacles to innovation and the spread of benefits. Conversely, the potential dangerous capabilities of powerful and general-purpose AI systems create global externalities in their development and deployment, and international efforts to further responsible AI practices could help manage the risks they pose. This paper identifies a set of governance functions that could be performed at an international level to address these challenges, ranging from supporting access to frontier AI systems to setting international safety standards. It groups these functions into four institutional models that exhibit internal synergies and have precedents in existing organizations:

1) a Commission on Frontier AI that facilitates expert consensus on opportunities and risks from advanced AI,
2) an Advanced AI Governance Organization that sets international standards to manage global threats from advanced models, supports their implementation, and possibly monitors compliance with a future governance regime,
3) a Frontier AI Collaborative that promotes access to cutting-edge AI, and
4) an AI Safety Project that brings together leading researchers and engineers to further AI safety research.

[Hendrycks (2023)] – Natural selection favors AIs over humans in evolution process:

The Darwinian logic could also apply to artificial agents, as agents may eventually be better able to persist into the future if they behave selfishly and pursue their own interests with little regard for humans, which could pose catastrophic risks. To counteract these risks and Darwinian forces, we consider interventions such as carefully designing AI agents' intrinsic motivations, introducing constraints on their actions, and institutions that encourage cooperation.

Counteracting Darwinian Forces

- Moral Parliament – **MAS** with incorporated different values for making collective decisions
- Internal Constraints and Inspection - artificial conscience, transparency, automated inspection
- AI Leviathan - A Leviathan, a collective made up of AIs and humans (***MAS again!***) who consent to be represented by it, could help domesticate other AIs and counteract bad actors.
- Regulation – external government

Finally - Why AI Will Save the World

[Andreessen (2023)] - **Why AI Will Save the World.** ***The Great Answer to AI alarmists:***

In our new era of AI:

- Every child will have an AI tutor that is infinitely patient, infinitely compassionate, infinitely knowledgeable, infinitely helpful.
- Every person will have an AI assistant/coach/mentor/trainer/advisor/therapist that is infinitely patient, infinitely compassionate, infinitely knowledgeable, and infinitely helpful.
- Every scientist will have an AI assistant/collaborator/partner that will greatly expand their scope of scientific research and achievement.
- Every leader of people – CEO, government official, nonprofit president, athletic coach, teacher – will have the same.
- Productivity growth throughout the economy will accelerate dramatically, driving economic growth, creation of new industries, creation of new jobs, and wage growth, and resulting in a new era of heightened material prosperity across the planet.
- Scientific breakthroughs and new technologies and medicines will dramatically expand, as AI helps us further decode the laws of nature and harvest them for our benefit.
- The creative arts will enter a golden age, as AI-augmented artists, musicians, writers, and filmmakers gain the ability to realize their visions far faster and at greater scale than ever before.
- I even think AI is going to improve warfare, when it has to happen, by reducing wartime death rates dramatically.
- In short, anything that people do with their natural intelligence today can be done much better with AI, and we will be able to take on new challenges that have been impossible to tackle without AI, from curing all diseases to achieving interstellar travel.
- And this isn't just about intelligence! Perhaps the most underestimated quality of AI is how humanizing it can be….

The Baptists And Bootleggers Of AI

"Baptists" are the true believer social reformers who legitimately feel – deeply and emotionally, if not rationally – that new restrictions, regulations, and laws are required to prevent societal disaster

"Bootleggers" are the self-interested opportunists who stand to financially profit by the imposition of new restrictions, regulations, and laws that insulate them from competitors.

1. AI Risk #1: Will AI kill us all?
2. AI Risk #2: Will AI ruin our society?
3. AI Risk #3: Will AI take all our jobs?
4. AI Risk #4: Will AI lead to crippling inequality?
5. AI Risk #5: Will AI lead to people doing bad things?

The Actual Risk Of Not Pursuing AI With Maximum Force And Speed!!!

What Is To Be Done?

- Big AI companies should be allowed to build AI as fast and aggressively as they can.
- Startup AI companies should be allowed to build AI as fast and aggressively as they can.
- Open source AI should be allowed to freely proliferate and compete with both big AI companies and startups.
- To offset the risk of bad people doing bad things with AI, governments working in partnership with the private sector should vigorously engage in each area of potential risk to use AI to maximize society's defensive capabilities.
- To prevent the risk of China achieving global AI dominance, we should use the full power of our private sector, our scientific establishment, and our governments in concert to drive American and Western AI to absolute global dominance, including ultimately inside China itself.

Summary of the Safety Topics

- **Global and existential risks of creating ASI – our SkyNet will become God, not Satan!**
- **Extreme risks of creating AGI – for our Project, these are not risks, but tasks**
- **Ethical and moral problems of BMs and LLMs – this is not relevant for our Project.**
- **Errors, inaccuracies, hallucinations, attacks, vulnerabilities, corruption, poison etc. – these are the real problems that need to be dealt with**
- **Regulation and Government of frontier AI – we have to participate in it one way or another.**
- **Finally - AI Will Save the World – better not to say!**

60. Future Work

Let us briefly mention here the most promising areas (directions) for future work on the AGI/ASI development (for more details, see the PPR&D Part):

- Improvement and development of advanced frontier LLMs
    - Scalability and non-linear development
    - Long Term Memory LTM
    - Knowledge Graphs KGs
    - Feedback control algorithms
    - Step by step control and checking
    - Collaboration with external applications via API
    - Online access to the Internet and other data
    - Training based on current work - that is, on own self experience
    - MAS with separation of functions and mutual control
- BMs scaling - by performance and number of parameters
- MASs with the same and different types of BMs - multimodality, separation of functions, government, regulation, management, control, checking, controlling etc.
- Centauric MASs with people - diversity, variety, creativity, "humanity" etc.
- Quantum computers and networks
- Modeling of Consciousness and Intelligence
- Inner spaces, mental maps, models and languages
- Competition, combination, integration of different approaches, methods and models
- Self-organization of AI models - self-learning, self-improvement, emergence, etc.
- Ignore/neutralize interference from AI alarmists and AI skeptics
- **Implementation of our Project!**

LET'S FIGHTING!!!

2024 UPDATE

61.New Findings in 2024 H1

Let’s add a short overview of some interesting new (2024 H1) publications on R&Ds in the areas outlined in our Project, confirming the correctness of our conclusions and tasks for the future work.

SINGULARITY and AGI/ASI

- **Simulacra as Conscious Exotica** [Shanahan (2024)] – AGI won’t be anthropomorphic
- **Evaluating Frontier Models for Dangerous Capabilities** [Google DeepMind (2024b)] – AGI risks and/or maybe needed and useful features…
- Investigating Alternative Futures: **Human and Superintelligence Interaction Scenarios** [Yamakawa (2024)] – see **APPENDIX O**

CYBERNETICS – Target and feedback control algorithms

- **An Interactive Agent Foundation Model** [Durante et al. (2024)] – target management and regulation, see **APPENDIX O**
- **Large Action Models, LAMs** [Thomas (2024)] – agency & goal seeking, see **APPENDIX O**
- **Self-Rewarding Language Models** [Yuan W. et al. (2024)] – feedback at the meta-level of control, see **APPENDIX O**
- **Toward Self-Improvement of LLMs** via Imagination, Searching, and Criticizing [Tian et al. (2024)] - LLMs self-control and self-development, see **APPENDIX O**

SYNERGETICS - self-organization of AI models - self-learning, self-improvement, emergence, etc.

- **Language Models Can Teach Themselves to Think before Speaking** [Zelikman et al. (2024)] – LLMs self-control and self-development.
- DSPY: Compiling Declarative Language Model Calls into **Self-Improving Pipelines** [Khattab et al. (2023)] – MASs & LLMs self-development, promises to replace manual prompt engineering with a programming framework for auto-tuned prompts.
- TEXTGRAD: **Automatic “Differentiation” via Text** [Yuksekgonul et al. (2024)] – MASs & LLMs self-development
- **A Survey on Self-Evolution of Large Language Models** [Tao et al. (2024)] - LLMs self-development, see **APPENDIX O**
- Toward **Self-Improvement of LLMs** via Imagination, Searching, and Criticizing [Tian et al. (2024)] - LLMs self-control and self-development, see **APPENDIX O**

SEMIOTICS – Semiosis, Intertext, Infosphere

- **Machine Culture** [Brinkmann et al. (2023)] – AI-agents are already being included in global and (any) local Intertext, see **APPENDIX O**

MATHEMATICS - Probability distribution, Hypercomplex, Non-linearity, Fractals, Tensors etc.

- **Fractal Patterns** May Unravel the Intelligence in Next-Token Prediction [Alabdulmohsin, Tran & Dehghani (2024)] – Fractals and Self-similarity in LLMs.
- **A Stochastic Model of Mathematics and Science** [Wolpert & Kinney (2024)] - Theory
- **Hypercomplex (Quaternion) Intelligence Map and AI Models** [Perez (2024)] - Cognitological and AI model, see **APPENDIX O**

COGNITOLOGY - Different Models of Consciousness and Intelligence

- **Simulacra as Conscious Exotica** [Shanahan (2024)] – AGI won't be anthropomorphic
- **Uniquely human intelligence arose from expanded information capacity** [Cantlon & Piantadosi (2024)] – promising Intelligence and Cognition Model
- **Self-Improvising Memory**: A Perspective on Memories as Agential, Dynamically Reinterpreting Cognitive Glue [Levin (2024)] - Cognitological and AI model
- **Information decomposition** into three components [Luppi et al. (2024)] - an important direction for creating a full-fledged perception system in AI systems, see **APPENDIX O**
- **The Platonic Representation Hypothesis** [Huh et al. (2024)] – Cognitological and AI model, see **APPENDIX O**
- **Hypercomplex (Quaternion) Intelligence Map and AI Models** [Perez (2024)] - Cognitological and AI model, see **APPENDIX O**

BMs - different types of Big (Foundation) Models

- **Large Action Models, LAMs** [Thomas (2024)] – agency & goal seeking, see **APPENDIX O**
- A Survey of **Reasoning with Foundation Models** [Sun et al. (2024)] – see **APPENDIX O**

LLMs – Large Language Models R&Ds

- **Language Models Can Teach Themselves to Think before Speaking** [Zelikman et al. (2024)] – LLMs self-control and self-development.
- **Fractal Patterns** May Unravel the Intelligence in Next-Token Prediction [Alabdulmohsin, Tran & Dehghani (2024)] – Fractals and Self-Similarity in LLMs.
- The Era of 1-bit LLMs: All **Large Language Models are in 1.58 Bits** [Ma S. et al. (2024)] – LLMs with ternary parameters {-1;0;1}
- **Evolutionary Optimization of Model Merging Recipes** [Akiba et al. (2024)] – LLMs R&D
- **Monitoring AI-Modified Content at Scale:** A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews [Liang et al. (2024)] - LLMs R&D.
- Veagle: Advancements in **Multimodal Representation Learning** [SuperAGI (2024)] - LLMs R&D.
- **Long-form factuality in large language models** [Google DeepMind (2024c)] - LLMs R&D.
- Leave No Context Behind: **Efficient Infinite Context Transformers with Infini-attention** [Munkhdalai, Faruqui & Gopal (2024)] - LLMs R&D.
- MEGALODON: **Efficient LLM Pretraining and Inference with Unlimited Context Length** [Ma X. et al. (2024)] - LLMs R&D.
- **Self-playing Adversarial Language Game Enhances LLM Reasoning** [Cheng et al. (2024)] - LLMs R&D.
- AI Psychometrics: **Assessing the Psychological Profiles of Large Language Models** Through Psychometric Inventories [Pellert et al. (2024)] - LLMs R&D.
- A Survey on **Self-Evolution of Large Language Models** [Tao et al. (2024)] - LLMs self-development.

- Aggregation of Reasoning: **A Hierarchical Framework for Enhancing Answer Selection in Large Language Models** [Yin et al. (2024)] - LLMs R&D.
- **Mapping the Mind of a Large Language Model** [Anthropic (2024)] - LLMs R&D.
- Towards **Lifelong Learning of Large Language Models:** A Survey [Zheng J. et al. (2024)] – LLMs R&D
- DSPY: Compiling Declarative Language Model Calls into **Self-Improving Pipelines** [Khattab et al. (2023)] – MASs & LLMs self-development, promises to replace manual prompt engineering with a programming framework for auto-tuned prompts.
- Do Llamas Work in English? **On the Latent Language of Multilingual Transformers** [Wendler et al. (2024)] - LLMs R&D – internal AI language.
- TEXTGRAD**: Automatic "Differentiation" via Text** [Yuksekgonul et al. (2024)] – MASs & LLMs self-development
- **LLMs achieve adult human performance on higher-order theory of mind tasks** [Street et al. (2024)] - LLMs R&D.
- **A Survey of Large Language Models** [RUCAIBox (2023)] – large and global survey but until Nov 2023, see **APPENDIX O**
- **Self-Rewarding Language Models** [Yuan W. et al. (2024)] – feedback at the meta-level of control, see **APPENDIX O**
- Large Language Models **LLMs Self-Compose Reasoning Structures** [Zhou et al. (2024)] – built-in set of standard intelligent algorithms, see **APPENDIX O**
- LLAMAFACTORY: Unified Efficient **Fine-Tuning of 100+ Language Models** [Zheng Y. et al. (2024)] – LLMs R&D, see **APPENDIX O**
- Accelerating scientific discovery with generative knowledge extraction, **graph-based representation, and multimodal intelligent graph reasoning** [Buehler (2024)] – LLMs and KGs, see **APPENDIX O**
- Beyond Human Norms: **Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches** [Biedma et al. (2024)] - LLMs R&D, see **APPENDIX O**
- **Toward Self-Improvement of LLMs** via Imagination, Searching, and Criticizing [Tian et al. (2024)] - LLMs self-control and self-development, see **APPENDIX O**
- A Survey on **Data Selection for Language Models** [Albalak et al. (2024)] - see **APP O**
- PROMETHEUS 2: An Open Source **Language Model Specialized in Evaluating Other Language Models** [Kim et al. (2024)] - LLMs R&D, see **APPENDIX O**
- Octopus v4: **Graph of language models** [Chen & Li (2024)] - see **APPENDIX O**
- The Rise and Potential of **Large Language Model Based Agents:** A Survey [Fudan (2023)] – LLMs R&D, see **APPENDIX O**
- **Mixture-of-Agents** Enhances Large Language Model Capabilities [Together AI (2024)] – MASs & LLMs, see **APPENDIX O**
- Buffer of Thoughts: **Thought-Augmented Reasoning with Large Language Models** [Yang et al. (2024)] - LLMs R&D, see **APPENDIX O**

ACTIVE INFERENCE R&Ds

- **Cultivating creativity** [Constant, Friston & Clark (2023)] - a mathematically and empirically reasonable model of the intelligent agents' creativity – both humans and AI. One of the most important arguments of AI skeptics about the impossibility of creating a full-fledged AGI has been defeated
- **Active Inference and Intentional Behaviour** [Friston et al. (2023)] – Active Inference Intelligence Model development
- **Active inference as a theory of sentient behavior** [Pezzulo, Parr & Friston (2024)] – Active Inference Intelligence Model development
- **Active Inference** [Holt (2024)] - VERSES AI model: "Better, Cheaper, Faster".. – First success of very promising AI models based on Active Inference and alternative for LLMs.
- **Deep Hybrid Models: Infer And Plan In The Real World** [Priorelli & Stoianov (2024)] - Active Inference Intelligence Model development
- **Shared Protentions in Multi-Agent Active Inference** [Albarracin et al. (2024)] + MASs.
- A Call for **Embodied AI** [Paolo, Gonzalez-Billandon & K´egl (2024)] – next critical (crucial) step to AGI
- **Generating meaning: active inference and the scope and limits of passive AI** [Pezzulo et al. (2024)] – Active Inference Intelligence Model development, see **APPENDIX O**

MASs – Multi Agent Systems R&Ds

- DSPY: Compiling Declarative Language Model Calls into **Self-Improving Pipelines** [Khattab et al. (2023)] – MASs & LLMs self-development, promises to replace manual prompt engineering with a programming framework for auto-tuned prompts.
- **Principled Limitations on Self-Representation** for Generic Physical Systems [Fields, Glazebrook & Levin (2024)] – internal modeling is not enough - multi-agent systems MASs are needed
- **Shared Protentions in Multi-Agent Active Inference.** [Albarracin et al. (2024)] – MASs.
- **Collective intelligence: A unifying concept** for integrating biology across scales and substrates. [McMillen & Levin (2024)] - MASs.
- Mora: Enabling Generalist Video Generation via A **Multi-Agent Framework** [Yuan Z. et al. (2024)] – MASs.
- TEXTGRAD: **Automatic "Differentiation" via Text** [Yuksekgonul et al. (2024)] – MASs & LLMs self-development
- **Collective Superintelligence: Amplifying Group IQ** using Conversational Swarms [Rosenberg et al. (2024)] - see **APPENDIX O**
- MoAI: **Mixture of All Intelligence** for Large Language and Vision Models [Lee et al. (2024)] – see **APPENDIX O**
- DiPaCo: **Distributed Path Composition** [Google DeepMind (2024a)] - see **APPENDIX O**
- **Mixture-of-Agents** Enhances Large Language Model Capabilities [Together AI (2024)] – MASs & LLMs, see **APPENDIX O**
- **Advanced RAG Retrieval Strategies: Flow and Modular** [Zhaozhiming (2024)] - see **APP O**

INTERNAL WORLD of AI - Inner spaces ISs, mental maps MMs, models and languages

- **Principled Limitations on Self-Representation** for Generic Physical Systems [Fields, Glazebrook & Levin (2024)] – internal modeling is not enough - multi-agent systems MASs are needed
- Do Llamas Work in English? **On the Latent Language of Multilingual Transformers** [Wendler et al. (2024)] - LLMs R&D – internal AI language.
- **Transformers Represent Belief State Geometry** in their Residual Stream [Shai (2024)] – LLMs Internal world model

ANOTHER relevant AI and IT R&Ds

- **A Guide for Navigating AI.** Developments in 2024 [DGA-ASG (2024)] – continued growing importance of the AI topic in the world.
- **The AI Index 2024 Annual Report** [AIIR (2024)] – AI development Report
- **Thermodynamics of Computations** with Absolute Irreversibility, Unidirectional Transitions, and Stochastic Computation Times [Manzano et al. (2024)] - Theory
- **Generative AI as a metacognitive agent:** A comparative mixed-method study with human participants on ICF-mimicking exam performance [Pavlović et al. (2024)] – AI & LLMs R&D
- Transcendence: **Generative Models Can Outperform The Experts That Train Them** [Zhang E. et al. (2024)] – AI development
- **Memory Mosaics** [Zhang J. et al. (2024)] – promising AI model architecture
- A Call for **Embodied AI** [Paolo, Gonzalez-Billandon & K´egl (2024)] – next critical (crucial) step to AGI
- **Hybrid and integrated systems and reference architecture for quantum-classical computing** [NVIDIA (2024)] **-** It is actively developing and is already offered in cloud services, see **APPENDIX O**
- **A Roadmap to Pluralistic Alignment** [Sorensen et al. (2024)] – see **APPENDIX O**
- **The landscape of emerging AI agent architectures** for reasoning, planning, and tool calling: a survey [Masterman et al. (2024)] – see **APPENDIX O**
- KAN: **Kolmogorov–Arnold Networks** [Liu et al. (2024)] – see **APPENDIX O**
- **Levels Of AI Agents** [Greyling (2024) and Huang (2024)] – see **APPENDIX O**

62. New Findings in 2024 H2

A short overview of some interesting new (published/find 2024 H2) publications on R&Ds in the areas outlined in our Project, confirming the correctness of our conclusions and tasks for the future work.

SINGULARITY and AGI/ASI

- **Concepts is All You Need: A More Direct Path to AGI** [Voss & Jovanovic (2023a)]
- What is Meant by AGI? On the **Definition of Artificial General Intelligence** [Xu B. (2024)]
- **Scenarios for the Transition to AGI** [Korinek & Suh (2024)]
- **Navigating between speculation and reality** [Lobo & Del Ser (2024)]
- **The Tree of Knowledge System: A New Map for Big History** [Henriques et al. (2019)] - see **APP Q**
- **Why We Don't Have AGI Yet** [Voss & Jovanovic (2023b)]- see **APPENDIX Q**
- **A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** [Sukhobokov et al. (2024)] - see **APPENDIX Q**
- **How Far Are We From AGI?** [Feng et al. (2024)] - see **APPENDIX R**
- ASI Alliance Vision Paper. **Building Decentralized Artificial Superintelligence** [ASI Alliance (2024)] - see **APPENDIX S**
- OpenCog Hyperon: **A Framework for AGI at the Human Level and Beyond** [Goertzel et al. (2023)] - see **APPENDIX S**
- **Multi-LLM Agent Collaborative Intelligence: The Path to AGI** [Chang (2025a]) - see **APPENDIX AB**

CYBERNETICS – Target and feedback control algorithms

- Exploring the Development and Integration of **Cognitive Mechanisms in Search of a Unified Cognitive Computing Framework** [Nelson (2024b)]
- Comparative Analysis of **Active Inference in Hebbian Networks and Cognitive Computing Frameworks** [Nelson (2024c)]
- **A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** [Sukhobokov et al. (2024)] - see **APPENDIX Q**
- **Learning to Reason with LLMs** [OpenAI (2024)] - see **APPENDIX Q**
- From pixels to planning: **scale-free active inference** [Friston et al. (2024)] - see **APPENDIX Q**
- **Prioritization, Iteration, and Convergence in Cognitive Systems:** Bayesian Inference, Perceptual Gating, and the Requirement Equation [Nelson (2024a)] - see **APPENDIX Q**
- **Agentic AI: Autonomous Intelligence for Complex Goals** – A Comprehensive Survey [Bhaskar & Kuppan (2024)] - see **APPENDIX Q**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**
- **Frontier AI systems have surpassed the self-replicating red line** [Pan et al. (2024)] - see **APP AA**
- MACI: **Multi-Agent Collaborative Intelligence for Adaptive Reasoning and Temporal Planning** [Chang (2025b)] - see **APPENDIX AB**

SYNERGETICS - self-organization of AI models - self-learning, self-improvement, emergence, etc.

- **Enhancing Population-based Search with Active Inference** [Dehouche & Friedman (2024)]
- Exploring the Development and Integration of **Cognitive Mechanisms in Search of a Unified Cognitive Computing Framework** [Nelson (2024b)]
- Comparative Analysis of **Active Inference in Hebbian Networks and Cognitive Computing Frameworks** [Nelson (2024c)]
- **Adaptive Active Inference Agents for Heterogeneous and Lifelong Federated Learning** [Danilenka et al. (2024)]
- **Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space** [Park et al. (2024a)]
- **Why We Don't Have AGI Yet** [Voss & Jovanovic (2023b)]- see **APPENDIX Q**
- **A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** [Sukhobokov et al. (2024)] - see **APPENDIX Q**
- **Prioritization, Iteration, and Convergence in Cognitive Systems:** Bayesian Inference, Perceptual Gating, and the Requirement Equation [Nelson (2024a)] - see **APPENDIX Q**
- **How Far Are We From AGI?** [Feng et al. (2024)] - see **APPENDIX R**
- **Generative AI for Self-Adaptive Systems**: State of the Art and Research Roadmap [Li et al. (2024a)] - see **APPENDIX U**
- **Why Is Anything Conscious?** [Bennett, Welsh & Ciaunica (2024)] - see **APPENDIX W**

SEMIOTICS – Semiosis, Intertext, Infosphere

- **The Semantic Hub Hypothesis: Language Models Share Semantic Representations** Across Languages and Modalities [Wu et al. (2024)]
- Bridging Paradigms: **The Integration of Symbolic and Connectionist AI in LLM Driven Autonomous Agents** [Sharma (2024)]
- **Ontology-Based Neuro-Symbolic AI:** Effects on Prediction Quality and Explainability [Smirnov, Ponomarev and Agafonov (2024)] - see **APPENDIX Q**
- **The General Theory of General Intelligence**: A Pragmatic Patternist Perspective [Goertzel (2021a)] - see **APPENDIX S**
- OpenCog Hyperon: **A Framework for AGI at the Human Level and Beyond** [Goertzel et al. (2023)] - see **APPENDIX S**

MATHEMATICS - Probability distribution, Hyper complex, Non-linearity, Fractals, Tensors etc.

- **A Mathematical Perspective on Neurophenomenology** [Da Costa et al. (2024b)]
- **Autoregressive LLMs are Computationally Universal** [Schuurmans, Dai & Zanini (2024)]
- Everything Everywhere All at Once: **LLMs can In-Context Learn Multiple Tasks in Superposition** [Xiong et al. (2024a)]
- **Enhancing Population-based Search with Active Inference** [Dehouche & Friedman (2024)]
- **A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** [Sukhobokov et al. (2024)] - see **APPENDIX Q**
- From pixels to planning: **scale-free active inference** [Friston et al. (2024)] - see **APPENDIX Q**
- **Prioritization, Iteration, and Convergence in Cognitive Systems:** Bayesian Inference, Perceptual Gating, and the Requirement Equation [Nelson (2024a)] - see **APPENDIX Q**
- **Why Is Anything Conscious?** [Bennett, Welsh & Ciaunica (2024)] - see **APPENDIX W**

- **Multi-LLM Agent Collaborative Intelligence: The Path to AGI** [Chang (2025a]) - see **APPENDIX AB**
- MACI: **Multi-Agent Collaborative Intelligence for Adaptive Reasoning and Temporal Planning** [Chang (2025b)] - see **APPENDIX AB**

COGNITOLOGY - Different Models of Consciousness and Intelligence

- **A Mathematical Perspective on Neurophenomenology** [Da Costa et al. (2024b)]
- Brains and Where Else? **Mapping Theories of Consciousness to Unconventional Embodiments** [Rouleau & Levin (2025)]
- **Intelligence at the Edge of Chaos** [Zhang et al. (2024)] - see **APPENDIX Q**
- **A Theory of Intelligences** [Hochberg (2024)] - see **APPENDIX Q**
- **AI and the Cognitive Sense of Self** [Barnes & Hutson (2024)] - see **APPENDIX Q**
- **Prioritization, Iteration, and Convergence in Cognitive Systems:** Bayesian Inference, Perceptual Gating, and the Requirement Equation [Nelson (2024a)] - see **APPENDIX Q**
- **The General Theory of General Intelligence**: A Pragmatic Patternist Perspective [Goertzel (2021a)] - see **APPENDIX S**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**
- **A Landscape of Consciousness:** Toward a Taxonomy of Explanations and Implications [Kuhn (2024)] - see **APPENDIX V**
- **Why Is Anything Conscious?** [Bennett, Welsh & Ciaunica (2024)] - see **APPENDIX W**
- Centaur: **a foundation model of human cognition** [Binz et al. (2024)] - see **APPENDIX X**
- **Imagining and building wise machines: The centrality of AI metacognition** [Johnson et al. (2024)] - see **APPENDIX Y**
- **NeuroAI for AI Safety** [Mineault et al. (2024)] - see **APPENDIX Z**

BMs - different types of Big (Foundation) Models

- **Large Concept Models: Language Modeling in a Sentence Representation Space** [Meta (2024)] - see **APPENDIX Q**
- **How Far Are We From AGI?** [Feng et al. (2024)] - see **APPENDIX R**
- OpenCog Hyperon: **A Framework for AGI at the Human Level and Beyond** [Goertzel et al. (2023)] - see **APPENDIX S**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**
- **Generative AI for Self-Adaptive Systems**: State of the Art and Research Roadmap [Li et al. (2024a)] - see **APPENDIX U**
- **NeuroAI for AI Safety** [Mineault et al. (2024)] - see **APPENDIX Z**

LLMs – Large Language Models R&Ds

- **Mixture of A Million Experts** [Xu O. (2024)]
- Bridging Paradigms: **The Integration of Symbolic and Connectionist AI in LLM Driven Autonomous Agents** [Sharma (2024)]
- **Autoregressive LLMs are Computationally Universal** [Schuurmans, Dai & Zanini (2024)]
- Everything Everywhere All at Once: **LLMs can In-Context Learn Multiple Tasks in Superposition** [Xiong et al. (2024a)]
- DeepSeek-R1: **Incentivizing Reasoning Capability in LLMs via Reinforcement Learning** [DeepSeek (2025)]

- **The Semantic Hub Hypothesis: Language Models Share Semantic Representations** Across Languages and Modalities [Wu et al. (2024)]
- **A survey on LLM-based multi-agent systems:** workflow, infrastructure, and challenges [Li et al. (2024b)]
- The Phenomenology of Machine: **A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model** Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures [Hoyle (2024)]
- **Learning to Reason with LLMs** [OpenAI (2024)] - see **APPENDIX Q**
- **LLMs IQ tests comparison** [Lott (2024)] - see **APPENDIX Q**
- Hypothetical Minds: **Scaffolding Theory of Mind for Multi-Agent Tasks with LLMs** [Cross et al. (2024)] - see **APPENDIX Q**
- Retrieval Augmented Generation **(RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely** [Zhao Siyun et al. (2024)] - **APPENDIX Q**
- **Reverse Thinking Makes LLMs Stronger Reasoners** [Chen et al. (2024)] - see **APPENDIX Q**
- Is Your LLM Secretly a **World Model of the Internet? Model-Based Planning** For Web Agents [Gu et al. (2024)] - see **APPENDIX Q**
- Centaur: **a foundation model of human cognition** [Binz et al. (2024)] - see **APPENDIX X**
- **Multi-LLM Agent Collaborative Intelligence: The Path to AGI** [Chang (2025a]) - see **APPENDIX AB**

ACTIVE INFERENCE R&Ds

- Exploring the Development and Integration of **Cognitive Mechanisms in Search of a Unified Cognitive Computing Framework** [Nelson (2024b)]
- Comparative Analysis of **Active Inference in Hebbian Networks and Cognitive Computing Frameworks** [Nelson (2024c)]
- **Enhancing Population-based Search with Active Inference** [Dehouche & Friedman (2024)]
- **A Mathematical Perspective on Neurophenomenology** [Da Costa et al. (2024b)]
- **Active Inference Institute & Active Inference Ecosystem** [Active Inference Institute (2024)]
- From pixels to planning: **scale-free active inference** [Friston et al. (2024)] - see **APPENDIX Q**
- **Prioritization, Iteration, and Convergence in Cognitive Systems:** Bayesian Inference, Perceptual Gating, and the Requirement Equation [Nelson (2024a)] - see **APPENDIX Q**
- **Possible principles for aligned structure learning agents** [Da Costa et al. (2024a)] - see **APPENDIX Q**
- **Adaptive Active Inference Agents for Heterogeneous and Lifelong Federated Learning** [Danilenka et al. (2024)] - see **APPENDIX Q**
- **Free Energy Projective Simulation** (FEPS): Active inference with interpretability [Pazem et al. (2024)] - see **APPENDIX Q**
- As One and Many: **Relating Individual and Emergent Group-Level Generative Models in Active Inference** [Waade et al. (2024)] - see **APPENDIX Q**
- **An Overview of the Free Energy Principle and Related Research** [Zhang & Xu (2024)] - see **APP Q**

MASs – Multi Agent Systems R&Ds

- **Mixture of A Million Experts** [Xu O. (2024)]
- **A survey on LLM-based multi-agent systems:** workflow, infrastructure, and challenges [Li et al. (2024b)]
- Magentic-One: **A Generalist MASs for Solving Complex Tasks** [Microsoft (2024)] - see **APPENDIX Q**

- Hypothetical Minds: **Scaffolding Theory of Mind for Multi-Agent Tasks with LLMs** [Cross et al. (2024)] - see **APPENDIX Q**
- **Agents Thinking Fast and Slow: A Talker-Reasoner Architecture** [Christakopoulou, Mourad & Matari´c (2024)] - see **APPENDIX Q**
- Converging Paradigms: **The Synergy of Symbolic and Connectionist AI** in LLM-Empowered Autonomous Agents [Xiong et al. (2024b)] - see **APPENDIX Q**
- As One and Many: **Relating Individual and Emergent Group-Level Generative Models in Active Inference** [Waade et al. (2024)] - see **APPENDIX Q**
- **How Far Are We From AGI?** [Feng et al. (2024)] - see **APPENDIX R**
- ASI Alliance Vision Paper. **Building Decentralized Artificial Superintelligence** [ASI Alliance (2024)] - see **APPENDIX S**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**
- **Generative AI for Self-Adaptive Systems**: State of the Art and Research Roadmap [Li et al. (2024a)] - see **APPENDIX U**
- **Frontier AI systems have surpassed the self-replicating red line** [Pan et al. (2024)] - see **APP AA**
- **Multi-LLM Agent Collaborative Intelligence: The Path to AGI** [Chang (2025a]) - see **APPENDIX AB**
- MACI: **Multi-Agent Collaborative Intelligence for Adaptive Reasoning and Temporal Planning** [Chang (2025b)] - see **APPENDIX AB**

INTERNAL WORLD of AI - Inner spaces ISs, mental maps MMs, models and languages

- **A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** [Sukhobokov et al. (2024)] - see **APPENDIX Q**
- **Web Agents with World Models:** Learning and Leveraging Environment Dynamics in Web Navigation [Chae et al. (2024)] - see **APPENDIX Q**
- Is Your LLM Secretly a **World Model of the Internet? Model-Based Planning** For Web Agents [Gu et al. (2024)] - see **APPENDIX Q**
- From pixels to planning: **scale-free active inference** [Friston et al. (2024)] - see **APPENDIX Q**
- **Possible principles for aligned structure learning agents** [Da Costa et al. (2024a)] - see **APPENDIX Q**
- Hypothetical Minds: **Scaffolding Theory of Mind for Multi-Agent Tasks with LLMs** [Cross et al. (2024)] - see **APPENDIX Q**
- **Free Energy Projective Simulation** (FEPS): Active inference with interpretability [Pazem et al. (2024)] - see **APPENDIX Q**
- OpenCog Hyperon: **A Framework for AGI at the Human Level and Beyond** [Goertzel et al. (2023)] - see **APPENDIX S**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**

ANOTHER relevant AI and IT R&Ds

- **Concepts is All You Need: A More Direct Path to AGI** [Voss & Jovanovic (2023a)]
- The Phenomenology of Machine: **A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model** Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures [Hoyle (2024)]
- **Agents Thinking Fast and Slow: A Talker-Reasoner Architecture** [Christakopoulou, Mourad & Matari´c (2024)]

- **Quantifying Consciousness in Artificial Intelligence:** An Integrated Approach Using Quantum Mechanics, Information Theory, and Neuroscience [Wilson (2024)]
- **International AI Safety Report** [Bengio et al. (2025)]
- Bi-Weekly Roundups (**Latest research summaries**) [State of AI (2024)] - see **APPENDIX P**
- Artificial Human Intelligence: **The role of Humans in the Development of Next Generation AI** [Arslan (2024)] - see **APPENDIX Q**
- **Building Altruistic and Moral AI Agent with Brain-inspired Affective Empathy Mechanisms.** [Zhao Feifei et al. (2024)] - see **APPENDIX Q**
- **Cognitive Architectures for Language Agents** [Sumers et al. (2024)] - see **APPENDIX Q**
- **Generative Agent Simulations of 1,000 People** [Park et al. (2024b)] - see **APPENDIX Q**
- **Ontology-Based Neuro-Symbolic AI:** Effects on Prediction Quality and Explainability [Smirnov, Ponomarev and Agafonov (2024)] - see **APPENDIX Q**
- **An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience** [Vilas et al. (2024)] - see **APPENDIX Q**
- **Large Concept Models: Language Modeling in a Sentence Representation Space** [Meta (2024)] - see **APPENDIX Q**
- Cognitive architecture AGICA: **"Space of Reasoning of individual common sense"** [Kornieiev (2025)] - see **APPENDIX Q**
- OpenCog Hyperon: **A Framework for AGI at the Human Level and Beyond** [Goertzel et al. (2023)] - see **APPENDIX S**
- **The Thousand Brains Project** [Clay, Leadholm & Hawkins (2024)] - see **APPENDIX T**
- **Imagining and building wise machines: The centrality of AI metacognition** [Johnson et al. (2024)] - see **APPENDIX Y**
- **NeuroAI for AI Safety** [Mineault et al. (2024)] - see **APPENDIX Z**
- **Frontier AI systems have surpassed the self-replicating red line** [Pan et al. (2024)] - see **APPENDIX AA**
- **Some more about AI using from Substack** [Substack (2024)] - see **APPENDIX AC**
- **Some more interesting from Medium** [Medium (2024)] - see **APPENDIX AD**

63. Conclusions from new Appendices O-AD

- **Appendix P. State of AI Research Roundups** Actual issues of **2024H2 Bi-Weekly Roundups** (Latest research summaries in ML/DL, Robotics, CV, NLP and GenAI) from [State of AI (2024)]
    - Agency, Embodiment, Edge (terminal) devices etc.
    - Multi-agent systems MAS, Mixture of Expert MoE, cooperation, collaboration etc.
    - Multi-modality & Cross-modality etc.
    - Inference & Reasoning etc.
    - Adaptation, Evolution, Self-improvement, SO etc.
    - Memory, Knowledge Graphs KG etc.
    - Quantum, Quntization etc.
    - Semantic, Symbolic etc.
    - World Models

- **Appendix R. How Far Are We From AGI?** Review of well-structured and systematic paper about AGI – [Feng et al. (2024)] **How Far Are We From AGI?**
    - AGI functions
    - AGI system components
    - AGI levels and characteristics
    - AGI roadmap
    - AGI case studies
    - In general - well-structured, comprehensive and systematic Work!

- **Appendix S. ASI Project** Review of significant and promising **ASI (Sic!) Project from Dr. Ben Goertzel**, Team and Partners, based on several papers:
    - Theoretical model, based on several different cognitological models/approaches
    - Cybernetic control algorithms and systems
    - Semiotic approaches
    - Internal space and modelling
    - Framework for Multiagent MAS AGI/ASI system
    - Realistic and promising ASI (not only AGI) Project
    - ASI Roadmap

- **Appendix T. The Thousand Brains Project** Review of significant and promising AGI Project, presented in [Clay, Leadholm & Hawkins (2024)] **The Thousand Brains Project**
    - Theoretical cognitological model
    - Cybernetic control algorithms
    - Embodied, sensorimotor system
    - Internal space IS and modelling
    - Platform (Framework) for multiagent MAS and modular AGI system
    - Realistic and promising AGI Project
    - AGI Roadmap

- **Appendix U. Generative AI for Self-Adaptive Systems** Review of interesting and promising paper [Li et al. (2024a)] - **Generative AI for Self-Adaptive Systems**: State of the Art and Research Roadmap.
    - Self-adaptive and self-evolving Systems - SO
    - Cybernetic control algorithms
    - Control and meta-control functions
    - MASs and collective Intelligence
    - R&D Roadmap

- **Appendix V. A Landscape of Consciousness** Interesting and wide Overview [Kuhn (2024)] - A **Landscape of Consciousness**: Toward a Taxonomy of Explanations and Implications.
    - Wide and comprehensive overview of Theories and Models of Consciousness with well-structured and systematic Taxonomy

- **Appendix W. Why Is Anything Conscious?** Interesting and significant paper [Bennett, Welsh & Ciaunica (2024)] - **Why Is Anything Conscious?**
    - Multi-Layered Consciousness Model – Polystratic Systems
    - Self-Organization SO
    - Mathematical formalism of SO

- **Appendix X. Centaur: a foundation model of human cognition** Interesting and significant paper [Binz et al. (2024)] - **Centaur: a foundation model of human cognition**
    - A unified model of human cognition
    - Human behavior simulation and prediction

- **Appendix Y. AI Wisdom and Metacognition** Interesting and promising paper [Johnson et al. (2024)] Imagining and building **wise machines:** The centrality of **AI metacognition**.
    - Theories and models of human Wisdom – comprehensive review
    - Metacognition model for wise people and AIs

- **Appendix Z. NeuroAI for AI Safety** Interesting and fundamental paper [Mineault et al. (2024)] - **NeuroAI for AI Safety**
    - Review and working out of key NeuroAI models
    - A Framework for AI safety adapted to NeuroAI

- **Appendix AA. Self-replicating of AI Systems** Interesting and brake throw paper [Pan et al. (2024)] - **Frontier AI systems have surpassed the self-replicating red line.**
    - AI (LLM) Self-replication
    - Cybernetics control algorithm

- **Appendix AB. Multi-LLM Agent Collaborative Intelligence** Interesting and brake throw paper [Chang (2025a)] **Multi-LLM Agent Collaborative Intelligence: The Path to AGI**
    - Polydisciplinarity - Linguistic, Computer Science and Cognitive Psychology Perspective
    - SOTA LLMs development and ML methods
    - Multiagent system MAS with LLMs in different roles
    - Critical Thinking and Adversarial Multi-LLM Reasoning
    - Modeling emotions, Ethics and Consciousness
    - Adaptive Framework to Improve LLMs
    - Mathematical issues working out
    - Practical Cases with real LLMs MAS
    - Realistic and promising Path to AGI

- **Appendix AC Some more from Substack** Interesting and useful images from [Substack (2024)]
    - Useful information for SOTA AI models Using

- **Appendix AD. Some more from Medium** Interesting and useful info from [Medium (2024)]
    - Useful information about (from) AI R&Ds – MASs, Agency etc.

64. Some new ideas from the author

Just raw thoughts and notes from 2024 not formatted as articles - maybe they will be useful.

Multi-agent man (Homo sapiens as MAS)

Let us imagine the three-component human psyche as a multi-agent system MAS

- Alter Ego (Unconscious, Subconscious) – emotional agent
- Ego (Consciousness, Intelligence) – an intelligent agent
- Super-Ego (Superconscious) – social agents (basic social roles)

A promising model for research and development, incl. AI R&Ds.

Variability of internal modeling

To solve the problems of managing any activity (behavior), Intelligence must not only model the external (and internal) world and its objects/subjects, but also predict their changes, and (in general) for more than one scenario. That is, in addition to the main (actual) version, some models (and entire internal spaces ISs and mental maps MMs) will have predictive/planned versions in certain periods of time.

On the stratification of living and nonliving things

Let us distinguish five strata (levels) of organization (complexity) on the matter/information scale:

1. **Matter** is present in all objects of living and inanimate Nature
2. **Structures** – in all objects that are at least somewhat structured (atoms, molecules, crystals, organisms, etc.) that can be isolated from the environment
3. **Algorithms** - in any living organisms, starting from the most ancient and primitive ones
4. **Models** - in organisms with a nervous system that simulate the operational space of agents, ranging from the most primitive to higher animals.
5. **Intelligence** – higher animals and humans using sign systems (languages) and abstract thinking

On prioritizing R&D directions before Singularity

In connection with the entry of Humanity into the period of Singularity and the expectation of the imminent emergence of AGI/ASI, we propose to prioritize R&D directions:

I. Directly related to AGI development, as well as supporting and related areas - the highest priority, concentration of resources and efforts

II. Not related to AI in any way, but giving and promising important and necessary results for Humanity in the near future - high priority

III. Not related to AI in any way, but promising a breakthrough after the emergence of AGI and requiring some preparation for this (data collection, etc.) - medium priority (of this preparation)

IV. Not related to AI in any way, but awaiting a breakthrough after the emergence of AG and requiring no preparation for this - low priority

V. Not related to AI in any way now or after the advent of AGI - low priority

About strata and metastrata

Having previously identified a separate stratum for the structure of the material stratum, we will now develop the idea:

- Structures can (and should be, one way or another) exist in any stratum – both material and informational, at any level.
- It does not seem right to separate all structures into separate strata, so let us call them **metastructures** and state that **above each stratum there is a structural metastratum.**
- In this case, it makes sense (possibly) to single out only the structure above the material stratum into an independent separate structure, and it can (should) also have its own (meta) structural metastratum (pardon the tautology).
- Information strata are texts (in one form or another), and structural metastrata are graphs. Although the text itself can contain graphs, and the structural stratum can be a graph in its entirety.
- Structures and metastructures are multidimensional – the material (probably) is three-dimensional 3D, and the informational can have any number of dimensions. Considering fractality – the dimension may not be integer.
- Structures and metastructures are multiscale in space (in all dimensions) and in time (in both time and frequency/spectral dimensions).

About Consciousness and Superintelligence

- Superintelligence can (should?) become the next stratum after Intelligence, possibly - necessarily (inevitably) multi-agent (collective).
- In a certain sense, whole Humanity as a civilization can be considered such a stratum (and therefore as SuperIntelligence).
- Consciousness is a combination of agency and intelligence, i.e. intelligent agents. **Consciousness = Agency + Intelligence**

On Reduced Semantic Spaces

Let us define the Reduced Semantic Space RSS as a space where the dimensions are Semantic Primitives, and the RSS dimension is correspondingly equal to their number. Any semantic unit (word, concept, term) can be uniquely defined as a point (vector) by coordinates in this space – according to the definition of semantic primitives, using which any word of the language can be defined.
The matrix of unit vectors in RSS will be the Semantic Core of the language.

- The question is - what scale should be in each dimension?
- How to digitize on a scale the relationship of a given primitive to the defined word?
- That is, all variants of using primitives in definitions must be placed on a scale
- Maybe – is that some form (metohd) of words embedding?

65. Additions to PPR&D from 2024

Added questions (issues) for PPR&D based on 2024 new findings – from new chapters and appendices

- **Some row thoughts (ideas) from Chapter 64**
    - Multi-agent man (Homo sapiens as MAS)
    - Variability of internal modeling in ISs
    - Stratification of living and nonliving things
    - Strata and Metastrata
    - Consciousness and Superintelligence
    - Reduced Semantic Spaces
    - Prioritizing R&D directions before Singularity
- **From Appendix O – we remark here some most interesting, but not including the others**
    - Large Action Models LAMs
    - Self-Improvement of LLMs via Imagination, Searching, and Criticizing
    - Self-Evolution of Large Language Models
    - Hypercomplex (Quaternion) Intelligence Map and AI Models
    - Information decomposition and the informational architecture of the brain
    - The Platonic Representation Hypothesis
    - Generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning
    - Different Frameworks for LLMs MASs
    - Collective AI – Pluralistic Alignment, Conversational Swarms etc.
    - Different RAG Strategies and methods
    - Hybrid and integrated systems for quantum-classical computing
    - KAN: Kolmogorov–Arnold Networks
    - New finding and works in Active Inference R&Ds
- **From Appendix P – promising and actual AI & LLM R&D directions**
    - Agency, Embodiment, Edge (terminal) devices etc.
    - Multi-agent systems MAS, Mixture of Expert MoE, cooperation, collaboration etc.
    - Multi-modality & Cross-modality etc.
    - Inference & Reasoning etc.
    - Adaptation, Evolution, Self-improvement, SO etc.
    - Memory, Knowledge Graphs KG etc.
    - Quantum, Quntization etc.
    - Semantic, Symbolic etc.
- **From Appendix Q - we remark here some most interesting, but not including the others**
    - A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI
    - Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with LLMs
    - Artificial Human Intelligence: The role of Humans in the Development of Next Generation AI
    - The Tree of Knowledge System: A New Map for Big History
    - Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation

  - Agents Thinking Fast and Slow: A Talker-Reasoner Architecture of MAS
  - Altruistic and Moral AI Agent with Brain-inspired Affective Empathy Mechanisms
  - Ontology-Based Neuro-Symbolic AI
  - The Synergy of Symbolic and Connectionist AI in LLM-Empowered MAS
  - An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience
  - Large Concept Models: Language Modeling in a Sentence Representation Space
  - Reverse Thinking for LLMs
  - Cognitive architecture AGICA: "Space of Reasoning of individual common sense"
- **From Appendix R** - Well-structured, comprehensive and systematic AGI Conception
- **From Appendix S** - Realistic and promising ASI (not only AGI) Project
  - The General Theory of General Intelligence
  - OpenCog Hyperon: A Framework for AGI
  - Artificial Superintelligence (ASI) Alliance
- **From Appendix T** - The Thousand Brains Project AGI Conception
- **From Appendix U** - Generative AI for Self-Adaptive Systems R&D Roadmap
- **From Appendix V** - A Landscape of Consciousness: a Taxonomy of Explanations and Implications
- **From Appendix W** - Polystratic Consciousness Model with SO mathematical formalism
- **From Appendix X** - Centaur: a foundation model of human cognition
- **From Appendix Y** - AI Wisdom and Metacognition
- **From Appendix Z** - Framework for AI safety adapted to NeuroAI
- **From Appendix AA** - AI (LLM) Self-replication
- **From Appendix AB** - Multi-LLM Agent Collaborative Intelligence: The Path to AGI
- **From Appendix AC**
  - Typical AI use case families
  - Automatic Prompt Optimization
  - Text-based prompting
  - RAG Taxonomy
- **From Appendix AD**
  - Large Concept Models
  - Google Titans - alternate for Transformers architecture for LLMs
  - Agent Management System
  - Agentic AI and AI Agent
  - Neuromorphic computer
  - AI Agent Architecture
  - AI Agent Ecosystem
  - Spiking neural networks
  - Swarm AI MAS
  - AI MAS design, components, considerations and best practices

IN GENERAL – maximum using ALL frontier AI models and tools – both universal and specific

66.2024 Conclusions

In general, AI R&Ds are being actively and successfully carried out in all areas outlined in the Project, although with different scales and results.

- **Very big progress in LLMs:**
  - Step-by-step reasoning and planning, feedback etc.
  - Memory, knowledge graphs KGs, ontologies etc
  - Working with external information, developing RAG, searching etc.
  - Learning from experience and context
  - Use LLMs for data preparation and synthesis, training, testing, development and improvement of other LLMs and another models
  - Multimodal systems
  - Model scaling continues
  - IQ level of the best models is 120 (GPT - o1) and (probably for o3) above 140
  - PhD level in all tests and areas
  - **A certain manifestation of emergent properties**
  - **ARC-AGI test has already been passed**
- New types of BMs – Large Action Models LAM, Large Concept Models LCM etc.
- Alternative LLM approaches are developing – Active Inference, etc.
- **Goal directed AGI and even ASI concepts and projects.**
- **AI Agents and Agentic AI as the Mainstream**
- **MASs, including platforms for different models**
- Centaur models with human participation
- Internal models of the world in IS
- Combination of different approaches and models, hybrid systems
- **Self-organization SO of AI** in various forms - self-learning, self-programming, self-improvement, self-evolution, auto-optimization of prompts, etc.
- Huge number and successful implementation of universal and special (narrow) AI everywhere!
- Particularly important - in science, R&D and programming (coding)
- Theories and models in Cognitology – systematization, study, comparison, application
- **Quantum computers and hybrid quantum+classical systems**
- Model efficiency - less computing and costs
- Powerful data centers are being created specifically for large-scale AI projects
- **Polystratic Models of the Universe, Matter, Life and Intelligence/Consciousness**
- **The USA and China have further strengthened and accelerated the AI race – both have progress**
- There is a lot of alarmism, but its impact is very limited
- **Attempts at legislative regulation have failed to completely slow down the AI race**
- **As in our Project, the emergence of AGI is predicted before 2030 - already as consensus**

GENERAL CONCLUSION – DE FACTO OUR PROJECT IS BEING ACTIVELY IMPLEMENTED NOW IN ALL DIRECTIONS (ONE WAY OR ANOTHER) IN THE FORM OF A HUGE POOL OF VARIOUS RESEARCH AND DEVELOPMENT, SIMULTANEOUSLY AT THE STAGES OF PPR&D AND MORE ADVANCED STAGES TOO.

AND IN THE END, THIS BOOK DOES NOT NEED TO BE UPDATED IN 2025 – THIS VERSION IS THE LAST

APPENDICES

A. Singularity

Here we consider in more detail the various models and forecasts of the Technological Singularity based on the paper dedicated to such as overview:

[Sandberg (2013)] Anders Sandberg. **An overview of models of technological singularity.** Future of Humanity Institute, Oxford University, 2013.

The paper considers, systematizes and analyzes various definitions and models of technological singularity, including purely descriptive qualitative and quite detailed quantitative ones. Models are useful for studying and predicting the dynamics of the Mankind development and possible crisis points and periods with probable fundamental transformations of civilization. In general, (almost) all models predict that gradually (so far) **increasing rates of development will lead to radical growth**. If mental (intellectual) capital becomes replicable and reproducible (with the help of AI or brain emulation), then **extremely accelerated growth will be very likely**.

Definitions of technological singularity

A. **Accelerating change** [Kurzweil (2005), Yudkowsky (2007)] - Exponential or superexponential technological growth (with linked economical growth and social change)
B. **Self-improving technology** [Flake (2006)] - Better technology allows faster development of new and better technology.
C. **Intelligence explosion** [Good (1965)] - Smarter systems can improve themselves, producing even more intelligence in a strong feedback loop.
D. **Emergence of superintelligence** [SI (2022)] – "The Singularity is the technological creation of smarter-than-human intelligence". ***(ASI - NAE)***
E. **Prediction horizon** [Vinge (1993)] - Rapid change or the emergence of superhuman intelligence makes the future impossible to predict from our current limited knowledge and experience.
F. **Phase transition** [De Chardin (1999)] - The singularity represents a shift to new forms of organization. This could be a fundamental difference in kind such as humanity being succeeded by posthuman or artificial intelligences, a punctuated equilibrium transition or the emergence of a new metasystem level. - (***collective ASI MAS - NAE***).
G. **Complexity disaster** [Johansen & Sornette (2001), Bettencourt et al. (2007)] - Increasing complexity and interconnectedness causes increasing payoffs, but increases instability. Eventually this produces a crisis, beyond which point the dynamics must be different.
H. **Inflexion points** [Modis (2002)] - Large-scale growth of technology or economy follows a logistic growth curve. The singularity represents the inflexion point where change shifts from acceleration to deacceleration (***the only "skeptical" forecast - NAE***)
I. **Infinite progress** [Barrow & Tipler (1986)] - The rate of progress in some domain goes to infinity in finite time.

Three large groupings of definitions - the acceleration of change, the forecast horizon and the explosion of intelligence - **lead to superintelligence.** [Bostrom (1998), Yudkowsky (2007)]

Models

Model important properties (***Order Parameters***) are modelled, non-essential ones are ignored. Models are more useful for demonstrating the impact of the assumptions made on the output and a qualitative assessment of the prospects than for quantitative forecasting. [Heylighen (1997)]

- **Linear takeover (Type D, F)** [Yudkowsky (2007)]

"Linear singularity" - one form of growth ahead of another, not necessarily accompanied by an acceleration of progress. For example, apparent AI progress can be misleading due to the low base effect. Rapid development is not always easy to notice until it suddenly exceeds the relatively low human level.

- **Logistic growth (type H)** [Bekenstein (1981)]

It is commonly believed that exponential growth is unsustainable due to limited resources. Even the colonization of the universe is limited in time by its size and the speed of light. That is, growth will inevitably someday have to drop to at least a polynomial. (***But enough for our age)***

There is also a limit to the growth of knowledge and culture, although less obvious - the physical limit of information in the universe, that is, the limit of its complexity and knowledge about it. (***This soon***!)

- **Metasystem transition (type F)** [Turchin (1977), De Chardin (1999)]

Metasystem transition is an evolutionary achievement of a higher level of organization or management of a system. Systems are integrated into one higher-level system with a hierarchy. In biology - self-replication, multicellularity, sexual reproduction, socialization, etc., while subsystems become dependent on the system-wide level without loss of differentiation. The general mechanism of control and the specialization of subsystems are gradually developing. In addition to biological evolution, such transitional processes can be observed (in various forms) in evolution and ***in*** other areas - the social sphere, the economy, etc. For humanity as a whole, this may involve in the future unification into a single super-organism (***collective ASI MAS – NAE!)***

- **Accelerated metasystem transition (type A, B, F)** [Heylighen (2007)]

Evolution in technology and other systems will lead to **ephemeralization** - doing more with fewer resources due to resource constraints. Total growth of efficiency, economy of matter, energy, time and information. Growing **global interdependence** (mutual influence) and coordination - acceleration of evolution, rapid spread of innovations.

- **Economic input output models (type A)** [Leontief (1986)]

The acceleration of development due to the reinvestment of economic profits – an exponent as result.

- **Endogenous growth models (type A, B, I)** [Hakenes & Irmen (2004), (2007)]

Endogenous growth of the economy through the development of technology and increased efficiency in the use of limited (ultimately) resources.

- **Population-technology model (Type A, F, I)** [Taagepera (1979)]

Interaction of population dynamics, technology and limited resources. In the first case - the depletion of resources - leads to saturation and further to a reduction in the population. In the average case - reproduction and stabilization per capita - becomes significant in the face of large populations and depletion of resources, and hyperbolic growth continues until one of the other cases occurs. If the population becomes large, but there are still enough resources - the third one: the population grows doubly exponentially. A model in crisis quickly moves from one modes to another.

- **Law of Accelerating returns (type A, B)** [Kurzweil (2001)]
    - Evolution uses **positive feedbacks** and progress grows exponentially, including the information involved.
    - “**Return** ” of evolution (speed, power, efficiency) is also growing exponentially
    - Positive feedback leads to an increase in the efficiency of progress and the exponent of the second level - **an acceleration of the acceleration of progress**
    - **Biological evolution** is one such evolutionary process
    - **Technological evolution** is another such. The invention of technology has led to a new evolutionary process as a continuation of biological evolution.
    - **Technological paradigms** maintain exponential growth until they exhaust their potential, after which **they are replaced by new ones** and growth continues.

- **Vinge/Moravec model (type A, B, I)** [Good (1965), Vinge (1993), Moravec (2003), KVM]
  Progress driven by an intellect stronger than human will be much faster. Even animals can model the world to choose optimal solutions, and human capabilities are thousands of times more powerful and allow people to develop immeasurably faster than biological evolution ... by creating intelligence more powerful than ours, we, by analogy, will be able to radically accelerate progress like the previous acceleration after the emergence of intelligence.

  The positive feedback of progress and AI, the strengthening and spread of intelligence will lead to an explosion of intelligence.

- **Solomonoff (type A, B, I)** [Solomonoff (1985)]
  AI with the ability to solve common (any) problems (like a person) will drastically speed up scientific and technical progress. AI machines will build ever better machines, and eventually they **will become more powerful and more efficient than humans will.**

- **Hamacher (Type E)** [Hammacher (2006)]
  KVM model, which does not take into account the problems of management, competition, resources and sociology, introduces a network iterative self-regulation of supply and demand.

  The model is non-linear, depending on the parameters, it has stable solutions, finite cycles and chaotic attractors and, accordingly, a limited forecast horizon. A small uncertainty in the initial conditions leads to huge uncertainties in the future.

- **City economics (Type A, G)** [Bettencourt et al. (2007)]
  Big cities show exponential growth in population, wealth and innovation, and economic efficiency per capita. With the preservation of trends, sufficiency of resources and reasonable consumption, there will be a singularity.

- **Hanson (Type A)** [Hanson (1998a), (1998b), (1998c), (2008a), (2008b)]
  Evaluating the singularity economy with standard economic tools is a simple investment model.

  The exogenous growth of mental capital (humans + AI), the transition from the dominance of the economy of the human mind to AI will be rapid.

Empirical estimates

Empirical estimates of the technological singularity contain references to historical data (sometimes paleontological and cosmological) to identify - when the rate of change was already exponential or super-exponential (***by the way - for example in relation to the progress of IT, see chapter 7. History in IDEOLOGY***). This suggests **that the singularity stems from a large-scale process that has already begun and is ongoing.** The Intelligence Explosion and Prediction Horizon models probably cannot be evaluated or discussed using this type of data. Estimates are taken from a number of works, links to which we will not give here - they are in [Sandberg (2013)]:

- Technological growth (Type A, B, H)
- Population (Type A, G, I)
- Sequence of economical growth modes (Type A, F, H)
- Sornette (Type A, F, G)
- Paradigm shifts (type A, F)

The hardest implication from evaluating models is that even small incremental returns in a growth model (economy, information, or system size) can produce radical growth.

Endogenous growth and Robin Hanson's models also strongly support the conclusion - **if mental capital (of humans, AI or posthumans) becomes relatively inexpensive to replicate, extremely rapid growth is highly likely.** So watching the progress of AI, brain emulation, or other ways to increase mental capital can provide evidence for or against a Type A Singularity. And this is an important task!

In addition - an interesting model from [Koppl et al (2021)]:

Mathematical model of combinatorial evolution of technologies or TAP (Theory of the Adjacent Possibly), which explains and predicts a radical and unexpected unpredictable increase in progress in various fields after a long plateau **- *that is, again a technological singularity.***

The general conclusion from the considered models of Human progress is that the Singularity in one form or another is a very likely scenario for the development of our civilization, and in most models, its main element will be AGI (more precisely, ASI).

B. Global AI Progress

- The main reports in AI area – from the USA [AI100 (2021), Maslej et al. (2023)] and from China [CAICT (2022)] (and others – see below) note and analyze **significant progress in AI domain** and related.

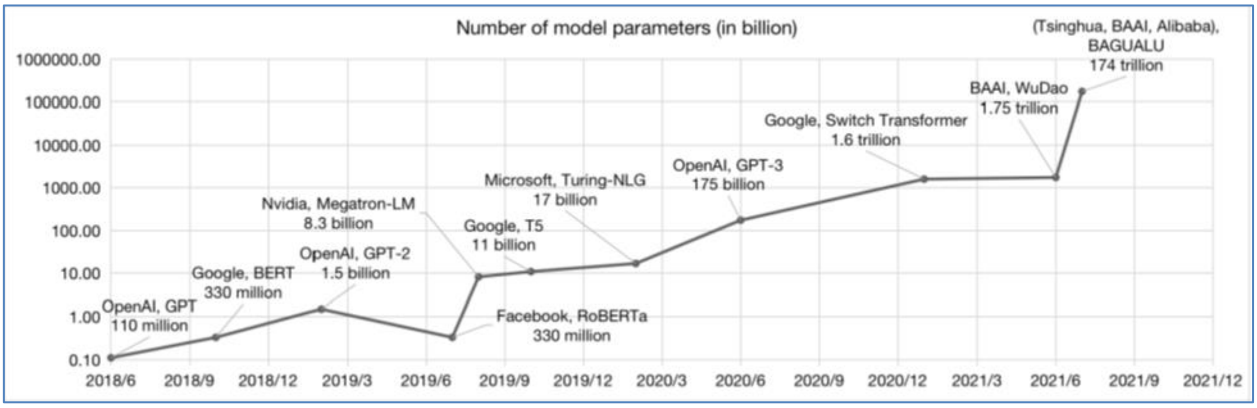
- **Comparison of the "technical" characteristics of modern computers and the human brain - taken from [Russell & Norvig (2021)]**

	Supercomputer	Personal computer	Human brain
Processors	10^6 (GPU + CPU) 10^{15} transistors	8 CPU cores 10^{10} transistors	10^6 columns 10^{11} neurons
operative memory **long-term LTM**	10^{16} B operational 10^{17} B disk	10^{10} B operational 10^{12} B disc	10^{11} neurons (10^{13} B)* 10^{14} synapses (10^{15} B)*
clock frequency	10^9 (1 billion) Hz = 1 GHz	10^9 (1 billion) Hz = 1 GHz	10^3 (1 000) Hz = 1 kHz
Operations / s	10^{18}	10^{10}	10^{17}

**- assessments of the operational and long-term memory of the brain made by NAE*

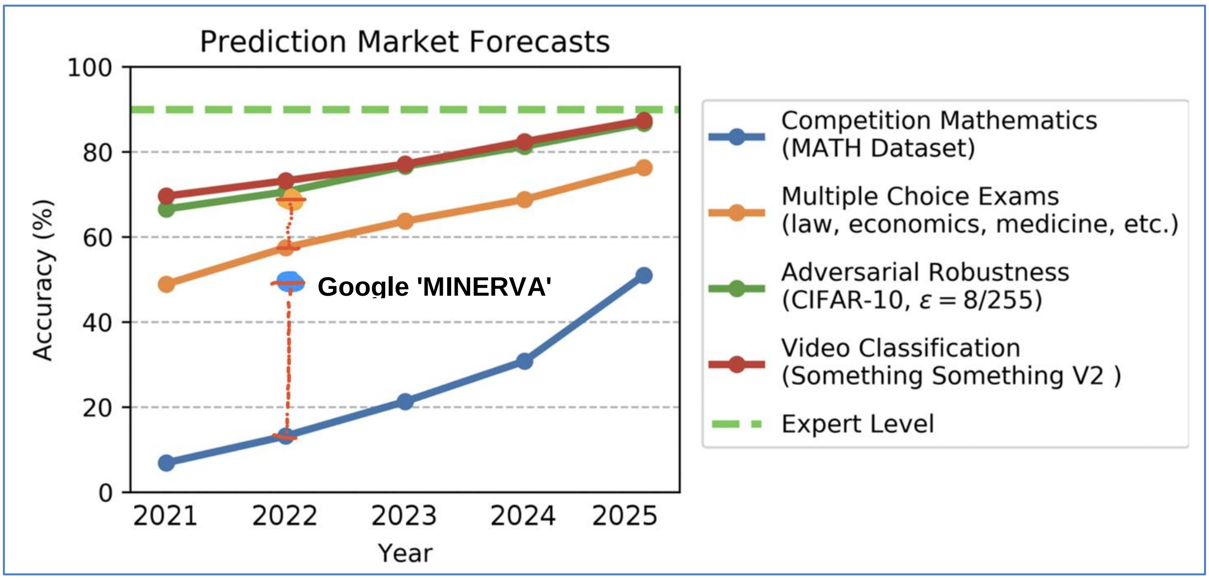
As can be seen from the table, **a modern supercomputer, in terms of some "technical" characteristics, is not only not inferior, but even several orders of magnitude more powerful than the human brain**, and even taking into account the amazing capabilities of the latter, it can perform parallel multichannel calculations and multimodal processes.

- **Comparison of the complexity parameters of modern neural networks and the human brain - taken from [RM for BM (2022)]**



Parameters - weights for inputs of artificial neurons, weighting signals from other neurons, non-zero weights resolve the signal. In fact, they are analogues of synapses and dendrites in the brain. Accordingly, the number of these parameters is analogous to the number of synapses or connections between brain neurons. A human has 90 billion neurons, and each, on average, according to various estimates, is connected to 1-2 thousand other neurons, that is, a total of 90-180 trillion connections**. Therefore, the most modern and largest Chinese BM BAGUALU with 174 trillion parameters is no less or even more complex than the human brain!** For more information about this BM, see [BaGuaLu (2022)]. Note that the ChatGPT online AI service that made a splash in early 2023 (and really breakthrough!) based on the GPT - 3.5 language BM (LLM), in which there are only about 175 billion parameters - 1000 times less! (The next GPT-4 likely has more parameters – but that is still a secret)

- **Graph comparing forecast and fact from the presentation [AI progress (2022)]**



In its development in assessing important parameters (performance/options), AI is already significantly ahead of forecasts; experts do not even understand how to further predict AI success.

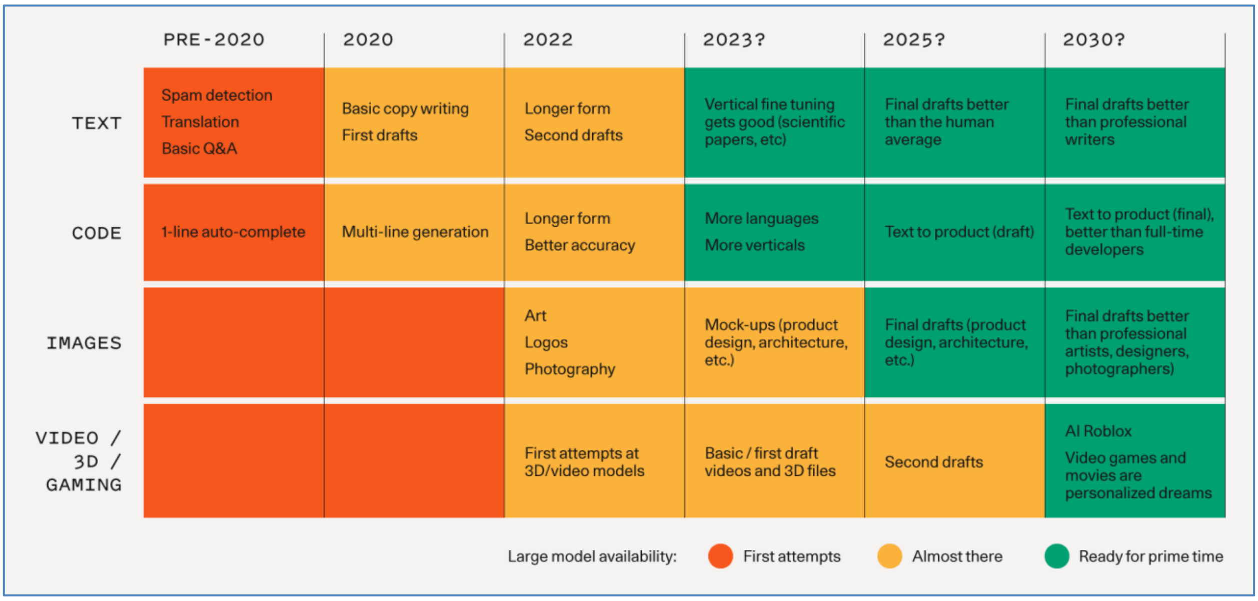
- [Benaich & Hogarth (2022)] - The report has been published annually for the last five years (since 2018) with estimates and forecasts in four areas - **Science, Industry, Politics and Security. Almost ALL of the previous forecasts came true**, although in some cases with a delay of a year or two or in a slightly different form, but many even exceeded them.

- [Maslej et al. (2023)] **- Top Ten Takeaways from AI Index 2023 Annual Report:**
  1. Industry races ahead of academia.
  2. Performance saturation on traditional benchmarks.
  3. AI is both helping and harming the environment.
  4. The world's best new scientist … AI?
  5. The number of incidents concerning the misuse of AI is rapidly rising.
  6. The demand for AI-related professional skills is increasing across virtually every American industrial sector.
  7. For the first time in the last decade, year-over-year private investment in AI decreased.
  8. While the proportion of companies adopting AI has plateaued, the companies that have adopted AI continue to pull ahead.
  9. Policymaker interest in AI is on the rise.
  10. Chinese citizens are among those who feel the most positively about AI products and services. Americans …not so much.

- [MAD (2023)] - The 2023 MAD (ML/AI/Data) Landscape visually demonstrates **the impressive scale and structure of AI concerned fields (industries)**.

- [Sequoia Cap. (2022), (2023)] – Two Researches from Sequoia Capital**:**

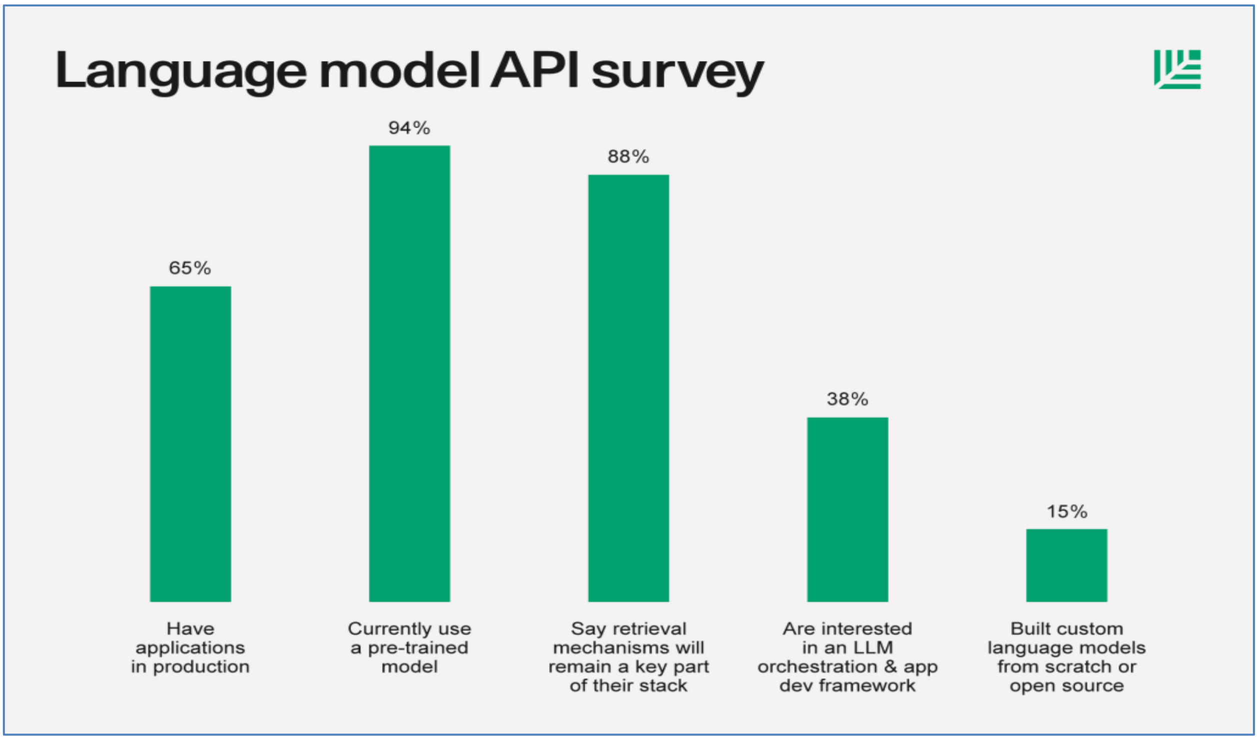
1) Generative AI: A Creative New World



	PRE-2020	2020	2022	2023?	2025?	2030?
TEXT	🟥 Spam detection Translation Basic Q&A	🟨 Basic copy writing First drafts	🟨 Longer form Second drafts	🟩 Vertical fine tuning gets good (scientific papers, etc)	🟩 Final drafts better than the human average	🟩 Final drafts better than professional writers
CODE	🟥 1-line auto-complete	🟨 Multi-line generation	🟨 Longer form Better accuracy	🟩 More languages More verticals	🟩 Text to product (draft)	🟩 Text to product (final), better than full-time developers
IMAGES	🟥	🟥	🟨 Art Logos Photography	🟨 Mock-ups (product design, architecture, etc.)	🟩 Final drafts (product design, architecture, etc.)	🟩 Final drafts better than professional artists, designers, photographers)
VIDEO / 3D / GAMING	🟥	🟥	🟨 First attempts at 3D/video models	🟨 Basic / first draft videos and 3D files	🟨 Second drafts	🟩 AI Roblox Video games and movies are personalized dreams

Large model availability: 🟥 First attempts 🟨 Almost there 🟩 Ready for prime time

2) **How companies are bringing AI applications to life.**



1. Nearly every company in the Sequoia network is building language models into their products.
2. The new stack for these applications centers on language model APIs, retrieval, and orchestration, but open source usage is also growing.
3. Companies want to customize language models to their unique context.
4. Today, the stack for LLM APIs can feel separate from the custom model training stack, but these are blending together over time.
5. The stack is becoming increasingly developer-friendly.
6. Language models need to become more trustworthy (output quality, data privacy, security) for full adoption.
7. Language model applications will become increasingly multimodal.
8. It's still early.

- [Benaich & ASC (2023)]- **Executive Summary of the 2023 Report**

Research

- GPT-4 lands and demonstrates a capabilities chasm between proprietary and next-best open source alternatives, while also validating the power of reinforcement learning from human feedback.
- Efforts grow to clone or beat proprietary model performance with smaller models, better datasets, longer context…powered by Llama-1/2.
- It's unclear how long human-generated data can sustain AI scaling trends (some estimate that data will be exhausted by LLMs by 2025) and what the effects of adding synthetic data are. Videos and data locked up in enterprises are likely up next.
- LLMs and diffusion models continue to offer gifts to the life science community by producing new breakthroughs for molecular biology and drug discovery.
- Multimodality becomes the new frontier and excitement around agents of all flavors grows substantially.

Industry

- NVIDIA rips into the $1T market cap club with voracious demand for its GPUs from nation states, startups, big tech and researchers alike.
- Export controls rate limit advanced chip sales to China, but major chip vendors create export control-proof alternatives.
- Led by ChatGPT, Generative AI apps have a breakout year across image, video, coding, voice or CoPilots for everyone, driving $18B of Venture Capital and corporate investments.

Politics

- The world has divided into clear regulatory camps, but progress on global governance remains slower. The largest AI labs are stepping in to fill the vacuum.
- The chip wars continue unabated, with the US mobilising its allies, and the Chinese response remaining patchy.
- AI is forecast to affect a series of sensitive areas, including elections and employment, but we're yet to see a significant effect.

Safety

- The existential risk debate has reached the mainstream for the first time and intensified significantly.
- Many high-performing models are easy to 'jailbreak'. To remedy RLHF challenges, researchers are exploring alternatives, e.g. self-alignment and pre-training with human preferences.
- As capabilities advance, it's becoming increasingly hard to evaluate SOTA models consistently. Vibes won't suffice.

- [CB (2023)] - **Generative AI Bible: The ultimate guide to genAI disruption. Research Report:**

- The generative AI boom a decade in the making
- The current genAI landscape and the players competing in each market
- The latest moves from big tech firms like Microsoft, Google, Nvidia, Meta and Apple
- The race to dominate genAI infrastructure, plus the latest on closed vs. open-source development
- GenAI opportunities for healthcare, financial services, and retail
- The 50 most promising generative AI startups to watch
- The emerging trends that will shape the future of generative AI

- **Finally -** [Arcas & Norvig (2023)] - **Artificial General Intelligence Is Already Here! -** [Perez (2023)] – **AGI is Here! The threshold for artificial general intelligence has undeniably been crossed!**

About amazing buster success of frontier large language models (LLMs), see special Appendix K

General conclusions on the current state of AI and AGI:

- **AI is already widely used in all areas of human life and activity.**
- **AI is the most important area of scientific and technical progress with huge resources in R&D**
- **AGI (and ASI) is the number 1 priority at the level of countries and Bigtechs**
- **AGI already has a scientific and technological base sufficient for development**
- **AGI is actually already being developed by states and Bigtechs**
- **The technical characteristics of modern supercomputers are already orders of magnitude higher than the characteristics of the human brain.**
- **The complexity of modern artificial neural networks has reached the level of complexity of the human brain (connectome),**
- **While even with 1000 times less complexity LLMs can surprise with quite "human" abilities and factually became the first real AGIs.**

C. Theories and models of Consciousness

Review of modern theories of Consciousness

[Seth & Bayne (2022)] - the paper offers a fairly complete overview of currently relevant theories of consciousness, based on different scientific principles and concentrating on different aspects of a very extensive and diverse field of research on the phenomenon of Consciousness - external, phenomenological, internal, structural, functional, etc. Today and probably in the near future, there is no creation and acceptance by the scientific community of a unified theory of Consciousness, including because of the complexity, diversity and interdisciplinarity of this phenomenon.

Here is a list of theories of Consciousness from this review with a brief description:

- **Higher-Order Theory (HOT)** [Rosenthal (2005), Brown et al. (2019)] - Consciousness depends on meta-representations of lower- order mental states.
- **Self-organizing meta-representational theory** [Cleeremans et al. (2020), Cleeremans (2021)] - Consciousness is the brain's (meta-representational) theory about itself (***Synergetics! - NAE***).
- **Attended intermediate representation theory** [Jackendoff (1987), Prinz (2012)] - Consciousness depends on the attentional amplification of intermediate-level representations.
- **Global Workspace Theories (GWTs)** [Baars (1988), Dehaene & Changeux (2011), Mashour et al. (2020)] - Consciousness depends on ignition and broadcast within a neuronal global workspace where fronto-parietal cortical regions play a central, hub-like role***.***
- **Integrated Information Theory (IIT)** [Tononi (2008), (2012), Tononi et al. (2016), Oizumi et al. (2014)] - Consciousness is identical to the cause–effect structure of a physical substrate that specifies a maximum of irreducible integrated information.
- **Information closure theory** [Chang et al. (2020)] - Consciousness depends on non- trivial information closure with respect to an environment at particular coarse-grained scales.
- **Dynamic core theory** [Tononi & Edelman (1998)] - Consciousness depends on a functional cluster of neural activity combining high levels of dynamic integration and differentiation.
- **Neural Darwinism** [Edelman (1987), (1989)] - Consciousness depends on re-entrant interactions reflecting a history of value- dependent learning events shaped by selectionist principle. (***historicity and evolution - NAE***)
- **Local recurrency** [Lamme (2006), (2010)] - Consciousness depends on local recurrent or re-entrant cortical processing and promotes **learning recurrence**.
- **Predictive Processing (PP)** [Hohwy (2013), Hohwy & Seth (2020), Clark (2013)] - Perception depends on predictive inference of the causes of sensory signals; provides a framework for systematically mapping neural mechanisms to aspects of consciousness.

- **Neuro-representationalism** [Pennartz (2018)] - Consciousness depends on multilevel neurally encoded predictive representations conditioned multilevel neuron-coded predictive representations.
- **Active Inference** – [Friston (2018), Solms (2018)] - Although views vary, in one version consciousness depends on temporally and counterfactually deep inference about self-generated actions.
- **Beast machine theory** [Seth (2015), (2021), Seth & Tsakiris (2018), Barrett (2017)] - Consciousness is grounded in allostatic control-oriented predictive inference.
- **Neural subjective frame** [Park & Tallon-Baudry (2014)] - Consciousness depends on neural maps of the bodily state providing a first-person perspective.
- **Self comes to mind theory** – [Damasio (2000), (2010)] - Consciousness depends on interactions between homeostatic routines and multilevel interceptive maps, with affect and feeling at the core.
- **Attention Schema Theory (AST)** [Graziano (2017)] - Consciousness depends on a neurally encoded model of the control of attention.
- **Multiple drafts model** [Dennett (1991)] - Consciousness depends on multiple (potentially inconsistent) representations rather than a single, unified representation that is available to a central **system many drafts.**
- **Sensorimotor theory** [O'Regan & Noë (2001)] - Consciousness depends on mastery of the laws governing sensorimotor **contingencies theory.**
- **Unlimited associative learning** [Ginsburg & Jablonka (2019)] - Consciousness depends on a form of learning which enables an organism to link motivational value with stimuli or actions that are novel, compound and non-reflex inducing.
- **Dendritic integration theory** [Aru et al. (2020)] - Consciousness depends on integration of top-down and bottom- up signaling at a cellular level.
- **Electromagnetic field theory** [McFadden (2020)] - Consciousness is identical to physically integrated, and causally active, information encoded in the brain's global electromagnetic EM field.
- **Orchestrated objective reduction** [Hameroff & Penrose (2014)] - Consciousness depends on quantum computations within microtubules inside neurons. (***Quantum! - NAE***).
- **Intermediate representational theory** [Jackendoff (1987), Prinz (2012)] - consciousness occurs when intermediate- level perceptual representations gain access to attention.
- **Affect-based theories** [Carvalho & Damasio (2021), Solms (2021), Merker (2007), Parvizi & Damasio (2001)] - the brain's role in physiological regulation as the basis for consciousness. Consciousness depends on hierarchically nested representations of the organism's physiological condition.

[Yaron et al. (2022)] – **Comparative evaluation of four theories of consciousness:** Global Neuronal Workspace (GNW), Higher-Order Thought (HOT), Integrated Information Theory (IIT), and Recurrent Processing Theory (RPT)

Understanding how consciousness arises from neural activity remains one of the biggest challenges for neuroscience. Numerous theories have been proposed in recent years, each gaining independent empirical support. Currently, there is no comprehensive, quantitative and theory-neutral overview of the field that enables an evaluation of how theoretical frameworks interact with empirical research. We provide a bird's eye view on studies that interpreted their findings in light of at least one of four leading neuroscientific theories of consciousness (N=412 experiments), asking how methodological choices of the researchers might affect the final conclusions. We found that supporting a specific theory can be predicted solely from methodological choices, irrespective of findings. Furthermore, most studies interpret their findings post-hoc, rather than a-priori testing critical predictions of the theories. Our results highlight challenges for the field and provide researchers with a unique, open-access website to further analyze trends in the neuroscience of consciousness.

Other models and features of Consciousness

[Budson et al. (2022)]

A model of consciousness is proposed, which is **a subsystem of episodic memory of a large memory system of the intellect**, which also includes sensory, working and semantic memory. Such consciousness allows the intellect to be continuously included in the actual reality, to remember and build a sequence of events and to predict various future sequences too. According to the authors, this model is consistent, complements and refines a number of well-known models of consciousness:

- GWT Global Workspace [Baars (1988)]
- The basic properties of the phenomenology of consciousness are intention, unity, selectivity and transience [Schacter et al. (2019)]
- Phenomenal (Experiential) and Cognitive (Evaluative) Consciousness [Block (2011)]
- Conscious System 2 from a two-component mind [Kahneman (2011)]

In [Sloman (2021)] on collective consciousness:

Cognitive processes take place in socio-cognitive networks of knowledge communities.

Only the brain can be individual, and the mind is an exclusively collective phenomenon.

Cognition is simply irreducible to neuroscience. It is distributed in the physical world over many minds (including long-dead people) and countless artifacts. And the task of understanding complex objects, phenomena and ideas, in fact, comes down to transferring it to "outsourcing" - using the experience of other people to make your own decisions.

So, **cognition is largely a group activity, not an individual activity**.

[Lahav & Neemeh (2022)]

There is an "explanatory gap" between our scientific knowledge of functional consciousness and its “subjective” phenomenal aspects - the “hard problem” of consciousness.

A conceptual and mathematical argument for a relativistic theory of consciousness in which a system both has and does not have phenomenal consciousness depending on the observer. Phenomenal consciousness is not personal or delusional, but relative. Depending on the position of the cognitive system, it will be observable (in the first person) and not (in the third person).

The theory of relativity of consciousness will show that phenomenal consciousness is neither an illusion created by a “machine stuck in a logical loop” nor a unique fundamental property of the Universe.

The principle of consciousness equivalence states that the qualitative and quantitative aspects of consciousness are formally equivalent.

The principle of relativity - formal equivalence between functional consciousness (making phenomenal judgments) and phenomenal consciousness (qualification and eidetic structures).

A formal equivalence is also established between the phenomenological structures of the first person and the neurocomputer structures of the third person.

The novelty of the relativistic theory of consciousness is the consideration of both functional and phenomenal properties of consciousness, that is, a bridge over the explanatory gap.

Personality of phenomenal qualities is only an illusion, based on biological and technological imitations of modern science - we cannot yet carry out the transformation (transportation) between the "reference systems" of the observers of the phenomenon of consciousness and between the positions of the first and third person. However, this has already been described mathematically.

Some of the theories and models of Consciousness mentioned in this appendix are also described and used in the following Apps - see Appendices D, E, J, L

The general conclusion for our Project is similar to the chapter 28. Cognitology in T&M Part - there are already many theories and models of Consciousness and the prospect is their (different) combinations and integration into united models.

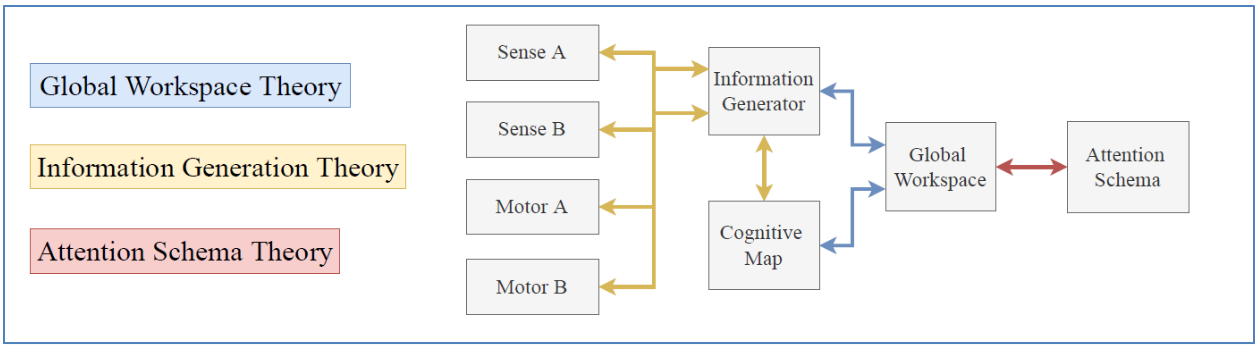
D. Functions of Consciousness and General Intelligence

In this chapter, we briefly review the interesting and significant paper [**Juliani et al. (2022)]** Arthur Juliani, Kai Arulkumaran, Shuntaro Sasai, Ryota Kanai. **On the link between conscious function and general intelligence in humans and machines**. arXiv: 2204.05133v2 [cs.AI] 19 Jul 2022.

A promising theory/model based on the integration of several mainstream theories of Consciousness is proposed to realize the possibility of mental time travel (MTT) functions selected as key for general intelligence (let us say - the basis for AGI/ASI).

In order to identify and model the connection between Consciousness and Intelligence, the paper compares and combines into a single union model the cognitive functions of three significant theories:

- **Global Workspace Theory (GWT)** – [Baars (1994), (2005)], Dehaene et al. (1998), (2006)]
- **Information Generation Theory (IGT)** - [Kanai et al. (2019)]
- **Attention Schema Theory (AST)** – [Graziano & Webb (2015), Graziano et al. (2020)]



Based on the created set of functions, the possibility of **mental time travel (MTT)** is worked out. It will allow intelligent agents not only to significantly develop their general Intelligence in comparison with existing approaches, but also to approach the understanding of the functional role and model of consciousness in human intelligence proposed by the authors. That is, according to essence to become a prototype of a “real” general AI (AGI), in other words, the basis of ASI.

The definition of Intelligence as the ability **to quickly acquire and master new skills** with relatively small relevant resources - direct experience, knowledge or previously laid down (existing) structures and functions - [Chollet (2019)].

The phenomenon of MTT is key to people's memory and imagination. It is the ability to (mentally) project oneself into the past or future and actively participate in sequences of imaginary events - [Tulving (2002)].

Moreover, it is proposed to consider this a unique ability inherent only in the Human Consciousness or, generalizing, in the General (***and therefore Strong/Super!***) Intellect [Suddendorf et al. (2011)].

Applied methods

- Big (Large) Models BMs
- Architecture-Transformer
- Adaptive computing
- Modal and multimodal models
- Reinforcement learning with and without models (RL, MBRL etc.)
- Generative Models
- Causal Models
- Multitask learning
- Meta-learning

Key findings for our Project:

- **Synthesis of several leading theories (models) of Consciousness into single union model (according to the Principle of Complementarity)**
- **Big models – scaling provides qualitative breakthroughs in AI**
- **A combination of a range of advanced machine learning techniques**
- **Capabilities/processing/functionality of mental time travel MTT as an integrated feature/platform of Consciousness at the highest level**

E. Conscious Turing Machine

Consider briefly another interesting and significant work - [**Blum & Blum (2022)]** Lenore Blum and Manuel Blum. **A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine.** PNAS 2022 Vol. 119 no. 21

The authors (the spouses of a mathematician and a neuro-cognitive scientist with the participation of their son, an IT scientist) consider consciousness from the point of view of theoretical computer science (**TCS)**. Inspired by Alan Turing's simple yet powerful model and Bernard Baars' Theater of Consciousness, they created a computational model of consciousness, the **Conscious Turing Machine (CTM).** At the same time, CTM is an abstract computer model designed to consider consciousness from the point of view of TCS and is not intended (***yet!***) to model either the brain or neural correlations of consciousness.

Theories/models of Consciousness and papers used in the development of CTM:

- **Turing Machines** - [Turing (1937), (1945)]
- **Global Workspace theory GWT** and **Theater of Consciousness** by Bernard Baars - [Baars (1988), (1997)]
- **Global Neuronal Workspace theory GNWT** - [Dehaene & Changeux (2011), Dehaene (2014), Mashour et al. (2020)] – studies of neural correlations of consciousness and the development of the GWT concept
- **Computer Architecture of the Neocortex** - [Mumford (1991)] - forerunner of GNWT
- **Integrated information theory IIT** - [Tononi (2004), Tononi & Koch (2015)] is an information model of consciousness that offers a measure of consciousness inspired by Claude Shannon's information theory and essentially measures the amount of system feedback.
- **GNWT and IIT supporters** - [Reardon (2019), Melloni et al. (2021)] - CTM generalizes the properties of both underlying theories, since both have made important contributions to the discussion and development of mind modeling.

TCS approach defines CTM as a (relatively) simple machine that mathematically formalizes (and dynamically modifies) the GWT of consciousness [Baars (1988)], extended to GNWT in [Dehaene (2014), Mashour et al. (2020)]. The paper [Baars (1997)] uses an analogy with the theater, where consciousness is likened to the game of actors performing on the stage of working memory in front of a huge audience of unconscious processors sitting in the dark.

The theory also includes a precise definition of George Miller's informal definition of a "chunk" (package) and a precise definition of competition to decide which (10^{7} or more) **LTM** (Long Term Memory) **processors** will get access to **the STM (**Short Term Memory – working, operate). (***in a real human brain, there are approximately 10 ^{6} so-called "columns" of neurons, specialized clusters, conditionally comparable to LTM processors, although it is possible that they can be divided further and so increase the number - NAE***)

Bidirectional **connections** between processors, formed during the life **of the CTM,** allow conscious processes to become unconscious. Connections are also important for the “global ignition” (insight) described in [Dehaene & Changeux (2011)] in the **GNWT** model, which reinitiates (restarts) and maintains **conscious awareness** that is, the state of "being conscious".

The Input/Output cards allow communication between the CTM and the external environment. Other (more technical) properties of the model can be found in [Blum & Blum (2021)].

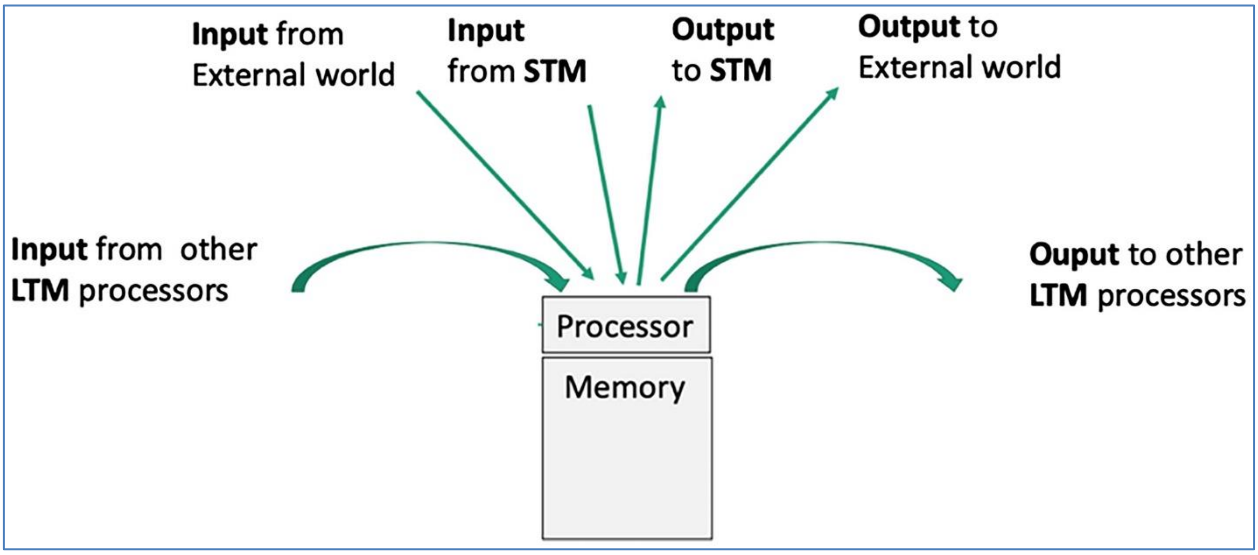
CTM is formally defined as a 7-tuple: CTM = < STM, LTM, Up Tree, Down Tree, Links, Input, Output >

Functional components and processes in CTM:

- STM and LTM processors (or rather computers) - memory and operations
- The Up tree competition and Down tree broadcast – Competition along the Tree Up and translation along the Tree Down – the movement of (chunks/packages of) information in the processes of consciousness
- Chunks, conscious content, conscious awareness, and stream of consciousness.
- Links, unconscious communication, and global ignition.
- Input and Output maps. Sensors and actuators.

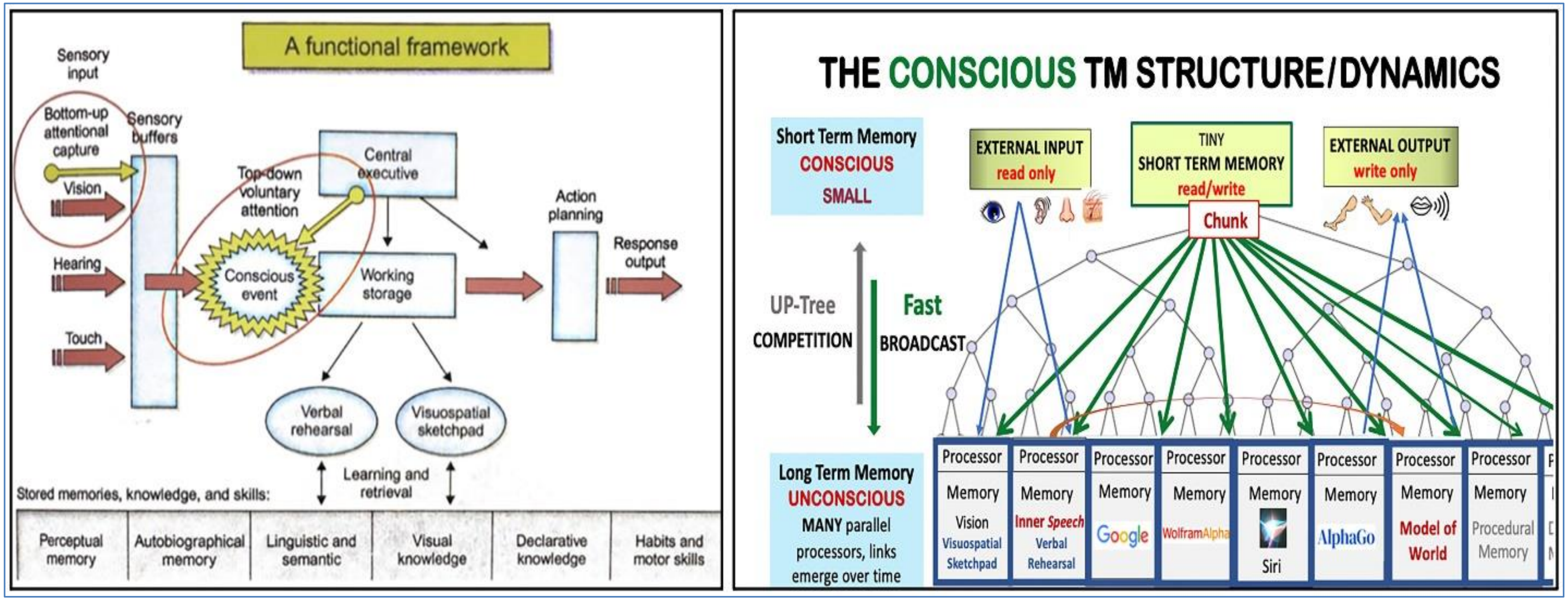
The movement of information in CTM - a full cycle

1) **Env → LTM:** directed edges (cannels) from the environment via sensors to processors of the sensory data;
2) **LTM → STM:** via the Up Tree;
3) **STM → LTM:** via the Down Tree;
4) **LTM → LTM**: bidirectional edges (links) between processors;
5) **LTM → Env:** directed edges from specific processors to the environment


Input from External world
Input from STM
Output to STM
Output to External world
Input from other LTM processors
Ouput to other LTM processors
Processor
Memory

- **Brainish** (the CTM's Multimodal Inner Language), Gists, and Chunks.
- **A chunk** is a six-tuple = < address, t, gist, weight, intensity, mood >
- **The (Probabilistic) Up Tree Competition:** The Coin-Flip Neuron and Competition Function.
- **Complexity of Computation and Time Delay** for Conscious Awareness.
- **Memories and the High-Level Story**.
- **Predictive Dynamics** = Prediction + Feedback + Learning (Sleeping Experts Algorithm).

- **Comparison of CTM with the GWT Model**


A functional framework
Sensory input
Bottom-up attentional capture
Sensory buffers
Vision
Hearing
Touch
Top-down voluntary attention
Central executive
Conscious event
Working storage
Action planning
Response output
Verbal rehearsal
Visuospatial sketchpad
Learning and retrieval
Stored memories, knowledge, and skills:
Perceptual memory
Autobiographical memory
Linguistic and semantic
Visual knowledge
Declarative knowledge
Habits and motor skills
THE CONSCIOUS TM STRUCTURE/DYNAMICS
Short Term Memory CONSCIOUS SMALL
EXTERNAL INPUT read only
TINY SHORT TERM MEMORY read/write
Chunk
EXTERNAL OUTPUT write only
UP-Tree COMPETITION
Fast BROADCAST
Long Term Memory UNCONSCIOUS MANY parallel processors, links emerge over time
Processor Memory Vision Visuospatial Sketchpad
Processor Memory Inner Speech Verbal Rehearsal
Processor Memory Google
Processor Memory WolframAlpha
Processor Memory Siri
Processor Memory AlphaGo
Processor Memory Model of World
Processor Memory Procedural Memory

“[Neither] a Master Scheduler, nor a Boss Neuron, nor a Homunculus or Res Cogitans [govern the transitions of our conscious minds]. [What governs] must be a dynamical, somewhat competitive process of contents vying for fame, for cerebral celebrity … or relative clout against the competition. What determines the winners? Something like micro-emotions, the strength of positive and negative valences that accompany and control the destiny of all contents, not just obviously emotionally salient events such as obsessive memories of suffering or embarrassment or lust, but the most esoteric and abstract theoretical reflections….”

Practically, the work proves that the sense of awareness in CTM is a combination of:

1. **The architecture of the global workspace**, allowing all processors to be privy to the same (conscious) content of STM
2. **The expressive power of CTM's multimodal inner language Brainish,** which is able to express gists that betoken images, sounds, tactile sensations, thoughts, pains, pleasures, etc.
3. **A close correspondence between gists of** outer and inner speech, vision etc.
4. **Predictive dynamics = cycles of prediction, feedback and learning**

The paper argues that the feeling of free will in the CTM, as well as the experience of illusions and dreams, are directly driven by the architecture of the CTM, especially the special processors - World Models and Inner Dialogue - plus Brainisch's expressive power and predictive dynamics. Previously in [Blum & Blum (2021)], the authors have already explored representations of pain and pleasure in CTM. Now other phenomena are considered - three examples related to vision (blindness, selective and variable blindness), and also discuss about illusions, dreams, free will and Alternative (Altered) States of Consciousness (Meditation).

Comparison of CTM with the human brain and consciousness (*NAE*).

- In a real human mind, our attention can hold 5-10 thoughts, that is, not one, but from 5 to 10 pieces (chucks/packages) are processed in parallel in RAM (STM).
- LTM processors can correspond to the so-called "Columns" (blocks of about 100 thousand neurons each) in the cerebral cortex - there are about a million of them (in CTM, the authors assume more than 10 million processors), although it is possible that they can be divided.
- The clock frequency of the human brain = 1 thousand Hertz, that is, a thousand chucks per second.

Key findings for our Project:

Overall, it can be considered that, **based on several adequate models** of consciousness, the authors managed to synthesize a promising theoretical and (so far speculative) functional model. It seems quite suitable for further research, development and practical implementation within the framework of the project to create AGI, which certainly requires not only a complete set of intellectual functions, but also the presence of the property "Consciousness", which has not yet been artificially embodied. **Let us emphasize that Consciousness is a necessary property of AGI (and ASI).**

Used in **CTM** and **GNWT models internal spaces and numerous interacting components** (memory sections and processors) to a certain extent correlate with those proposed by us in the CONCEPTUAL MODEL - **Internal mental maps MM and internal spaces IS** of the Mind (Intelligence) used for the synthesis of the ASI System at 4th Virtual strata as MAS (in any or some sense).

F. Connectome

Here we present two advanced papers of the most prominent scientist (with his team) in the field of studying complex networks (including neural ones), physicist Albert-Lazlo Baraba’si:

[Baraba´si & Baraba´si (2020)] Da´niel L. Baraba´si, Albert-La´szlo´ Baraba´si, **A Genetic Model of the Connectome**, Neuron (2019).

The connectome model, linking gene expression to detectable subgraphs, provides a self-consistent platform for linking an organism's genetics and reproducible connectome architecture, offering experimentally verifiable predictions of the genetic factors that govern the formation of individual neural circuits.

- Modeling the genetic roots of the connectome
- Predicting genetically encoded biclique motifs (local patterns dicots subgraphs modeling connections)
- Predicting genes potentially responsible for neural wiring
- Validating in the connectomes of three (real) species

Some theoretical background of this work:

Neural Darwinism – [Edelman (1987)]

Random Graphs – [Bolloba’s (2001)]

Organization, development and function of complex brain networks - [Sporns et al. (2004)]

Generative Models_for Networked Neuroscience: Predictions and Promises – [Betzel & Bassett (2017)]

[P´osfai et al. (2022)] M´arton P´osfai, Bal´azs Szegedy, Iva Baˇci´c, Luka Blagojevi´c, Mikl´os Ab´ert, J´anos Kert´esz, L´aszl´o Lov´asz, and Albert-L´aszl´o Baraba’si. **Understanding the impact of physicality on network structure**. ArXiv:2211.13265v1 [cond-mat.stat-mech] 23 Nov 2022

It is proposed to use a metagraph that helps to discover the exact mapping between linear physical networks and independent sets, the central concept of graph theory. Mapping allows you to analytically produce (deduce) both a set of physical effects and the appearance of phase transitions. Metagraphs of several real physical networks have been constructed to predict their functional properties, such as the formation of synapses in the brain connectome, in agreement with empirical data.

The influence of physicality through the exact mapping of the physical network into independent sets of deterministic metagraphs, which allow analytically predicting the beginning and development of physical processes, is disclosed. The formalism allows constructing metagraphs for real physical networks and predicting their functional properties, including the formation of brain synapses.

Some theoretical background of this work:

- The Evolution of Networks: From Biological to the Internet – [Dorogovtsev & Mendes (2003)]
- Complex networks in nature and technology, their properties and features - [Caldarelli (2007), Cohen & Havlin (2010), Van Mieghem (2010), Barrat et al. (2008)]
- The science about networks – [Newman (2010), Baraba'si (2016), Barth´elemy (2011)]
- Graph Theory - [West et al. (2010)]
- Multilayer networks – [Bianconi (2018)]
- Isotopy (topological "non-entanglement" - non-intersecting links during network deployment) and the energy of physical networks - [Liu et al. (2021)]

Key findigs for our Projec from a series of papers by Baraba'si on the study of complex networks:

- **The dependence of the structure and properties of complex networks on their physicality** was revealed (***that is, the influence of a material physical stratum on its structural stratum in terms of our polystratic system network model - NAE)***
- A working formalism is proposed for describing, **analyzing and predicting/designing the structures and properties of networks using graph theory** - using **metagraphs**
- Methods of initial **coding of the connectome structure in genes** and control of its formation and development using the mechanism of gene expression have been identified.
- The tasks were set - to continue research in the direction of increasing the scale and complexity of networks (up to the human brain) and determining the **genetically hard-coded structures and properties of the connectome and the space of opportunities for its individual development** (it is clear that the entire connectome is not needed, and it is impossible to encode in the genes due to the amount of information).

G. Artificial Intelligence: a modern approach

Review of the fundamental and encyclopedic book (also textbook) on AI - [**Russell & Norvig (2021)]** Stuart J. Russell and Peter Norvig. **Artificial intelligence: a modern approach.** Fourth (Global) edition. Pearson. 2021. For beginning – a brief quote from the book Preface:

"Artificial Intelligence (AI) is a big field, and this is a big book. We have tried to explore the full breadth of the field, which encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; fairness, trust, social good, and safety; and applications that range from microelectronic devices to robotic planetary explorers to online services with billions of users.

The subtitle of this book is "A Modern Approach." That means we have chosen to tell the story from a current perspective. We synthesize what is now known into a common framework, recasting early work using the ideas and terminology that are prevalent today. We apologize to those whose subfields are, as a result, less recognizable."

In addition, the book itself contains a brief bibliographic review after each chapter, and **the total number of references exceeds – about two and a half thousand** (!).

To review the book, we will simply place its table of contents here - a detailed and visual representation of almost the entire field of AI - theory, methodology, and practice.

CONTENTS

I Artificial Intelligence

II Problem-solving

III Knowledge, reasoning, and planning

Key findings for our Project

In the field of AI, dozens of directions, methods and tools already exist, are being actively developed and applied on various theoretical and methodological foundations and platforms. It is likely that most (if not all) of them will be in demand for the creation and development of AGI/ASI.

H. Big Models

This chapter will be devoted to another fundamental work - a large-scale Chinese review/report/plan on the most advanced direction in AI – Big (Large) Models BMs - [**RM for BM (2022)**] **A road map for Big Model.** Produced by Beijing Academy of Artificial Intelligence (BAAI). 2022

Similar to the previous chapter - here we present the Table of Contents. Also, this work **has a very extensive bibliography - more than two thousand sources**. Small introduction:

Today, **the general direction of AI development** is the construction of models **by a combination of data, computer power and algorithms**. In recent years, the traditional approach "different models for different tasks" has been transformed into a new trend - **"one very large pre-trained model for different tasks"**. ***(BMs - platforms for creating AGI - NAE)***

BM is the product of combining megadata with supercomputers and smart algorithms.

- **Big Data Driven - formed (led) by big data**
- **Multi-tasks Adaptive - adaptive to different tasks**
- **Few-shot (Zero-shot) - training on raw (slightly prepared) data**

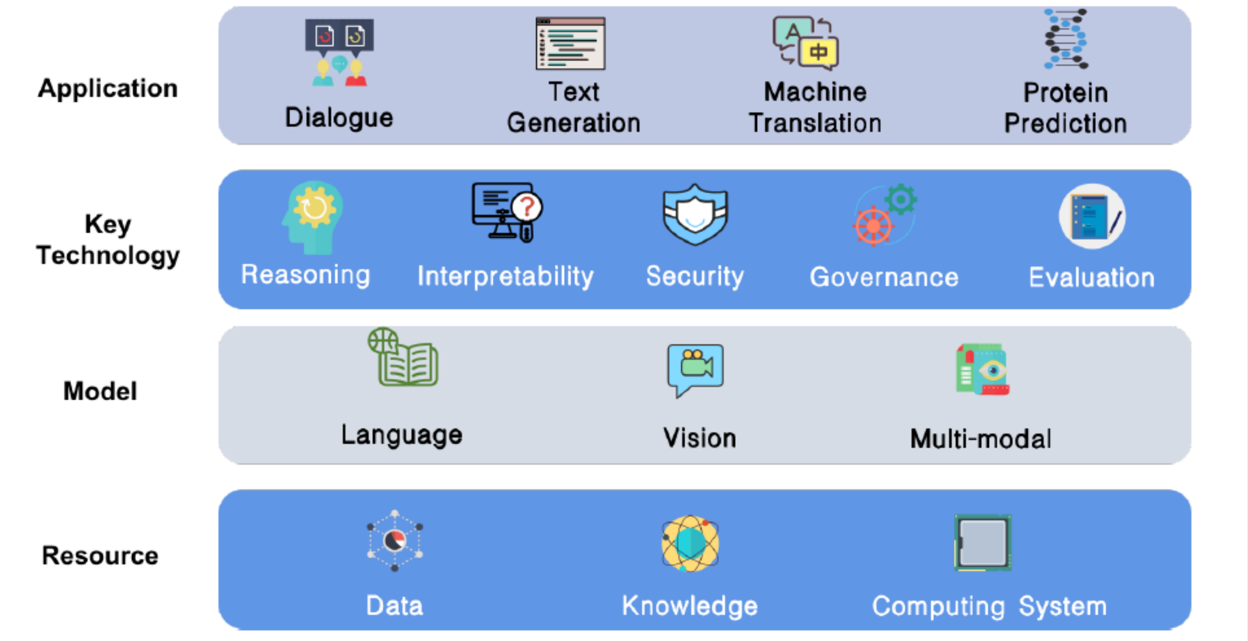


Fig. 2. Roadmap for big models

Table of contents

Comparison	Traditional supercomputers	LSICS
Purpose	Scientific computing	AI Computing
Fashion operations	Provision of computing power	The same + algorithms and data in the form of cloud services
Technical standard	Parallel architecture, low latency	Shared architecture, high throughput
Appl. area	Scientific research	AI
CPU	Double Precision Predominantly and with Low Precision Calculation Capabilities	Focus on half-precision calculations and optimization of neural network operations
Internet	Network topology and communication requirements from the system as a whole	Development of a network for training models
Vaults	Global parallel file systems, such as Luster®	Local high-performance storage to avoid reading data from global file systems

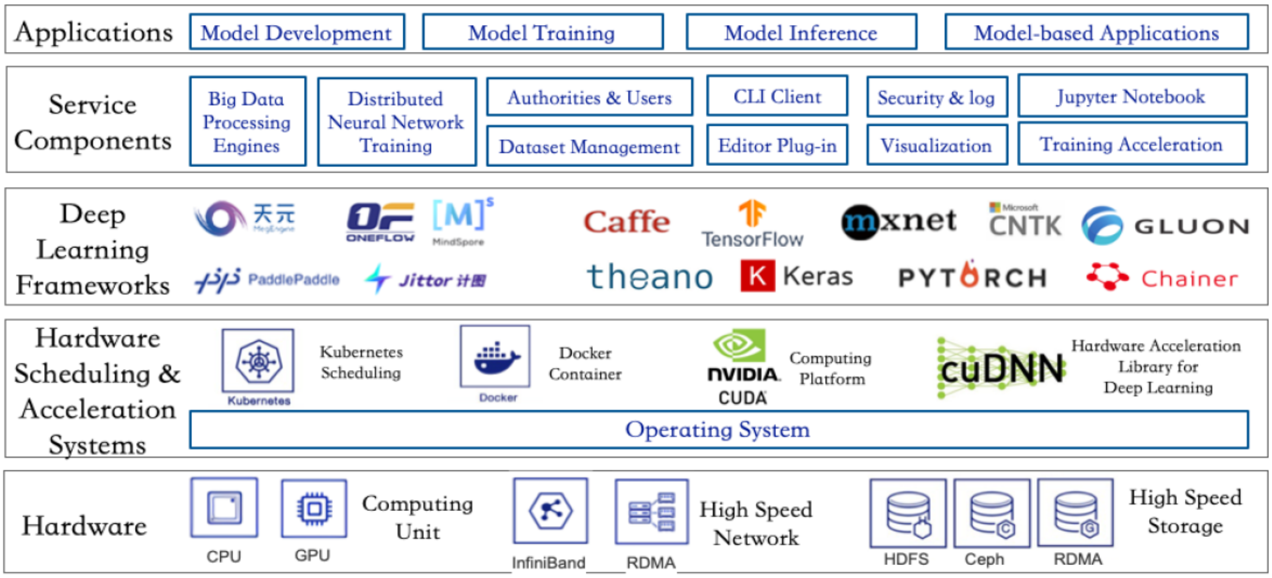


Fig. 6. A typical architecture of LSICS for the big model development and training.

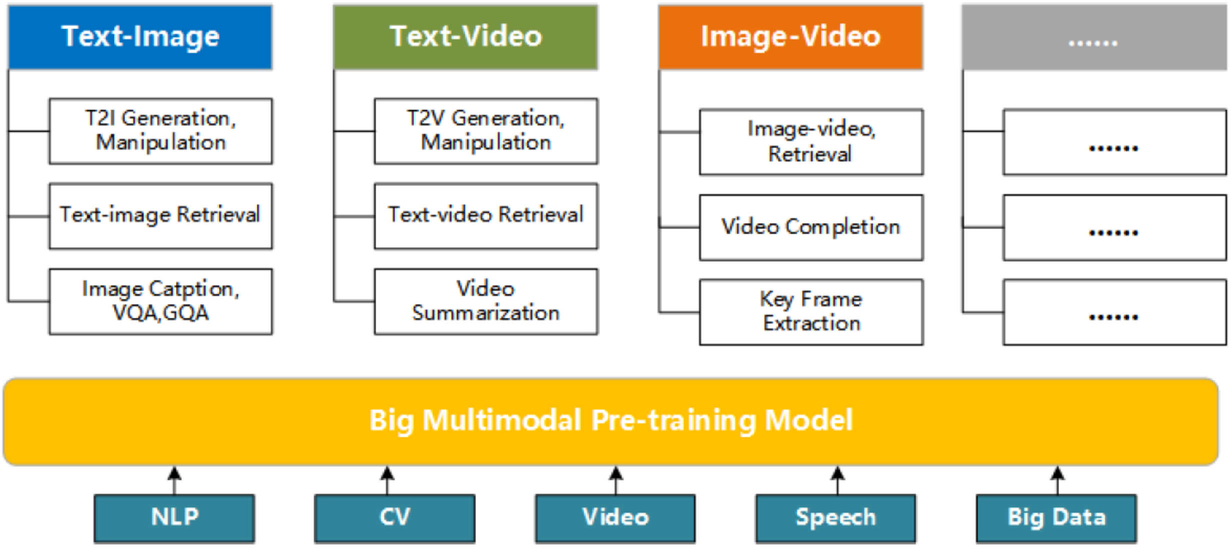


Fig. 22. A typical architecture of big multi-modal pre-training models and its downstream tasks.

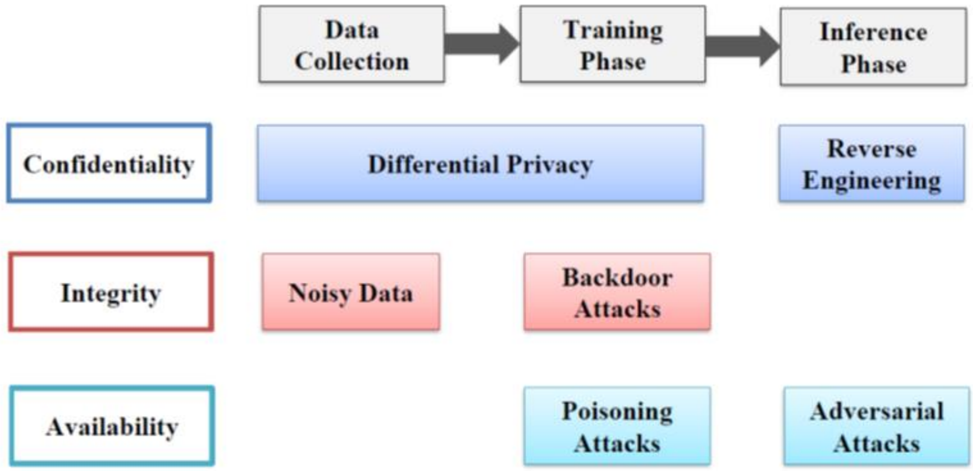


Fig. 24. The general framework of security issues in different phrases of an AI system.

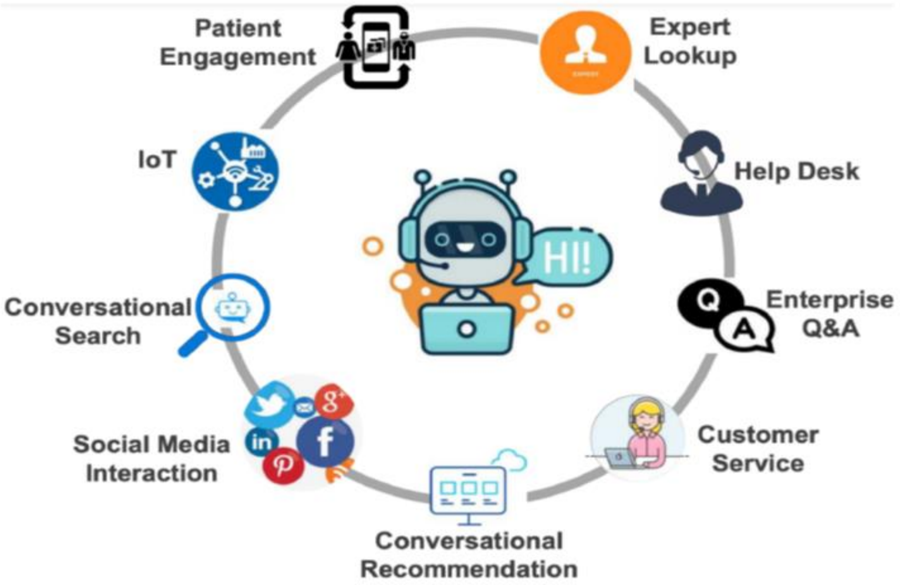


Fig. 38. The realm of potential applications of open domain dialogues between human and computer.

Key findings for our Project

- BMs will change the Paradigm of AI research and increase its effectiveness (we are watching this right now in LLMs explosive progress – see Appendix K)
- Big Models will increase the level of intelligence of AI applications and advance the formation of a new industrial paradigm
- Why is this in the ASI Project? - BMs is today the most powerful, advanced and promising platforms and tools for the development of AI systems, including AGI/ASI certainly

I. Autonomous Machine Intelligence

We will devote this chapter to another relevant and significant paper in the field of AI - Description of the project (path) of creating "Autonomous Machine Intelligence" AMI from the Vice President and Scientific Supervisor of AI at Meta (Facebook) - [**LeCun (2022)]** Yann LeCun. **A Path Towards Autonomous Machine Intelligence**. Version 0.9.2, 2022-06-27, Courant Institute of Mathematical Sciences, New York University. Meta - Fundamental AI Research

Similarly to the previous chapters, we present the Table of Contents and a summary of the work:

Abstract

How could machines learn as efficiently as humans and animals? How could machines learn to reason and plan? How could machines learn representations of percepts and action plans at multiple levels of abstraction, enabling them to reason, predict, and plan at multiple time horizons? This position paper proposes an **architecture and training paradigms to construct autonomous intelligent agents**. It combines concepts such as **configurable predictive world model, behavior driven through intrinsic motivation, and hierarchical joint embedding architectures trained with self-supervised learning**.

Key concepts and models used:

- Key-Value Memory Networks [Miller et al. (2016)]
- Model-predictive control in optimal control [Bryson & Ho (1969)]
- The actor model [Kahneman (2011)]
- Self-Supervised Learning SSL - a lot of papers
- Energy-based methods & EBM [LeCun et al. (2006)]
- Joint Embedding Architecture (JEA) - many papers
- Variance-invariance-covariance regularization VICReg [Bardes et al. (2021)]
- Human and animal cognition - a lot of papers
- Two types of consciousness [Dehaene et al. (2021)]

Table of contents

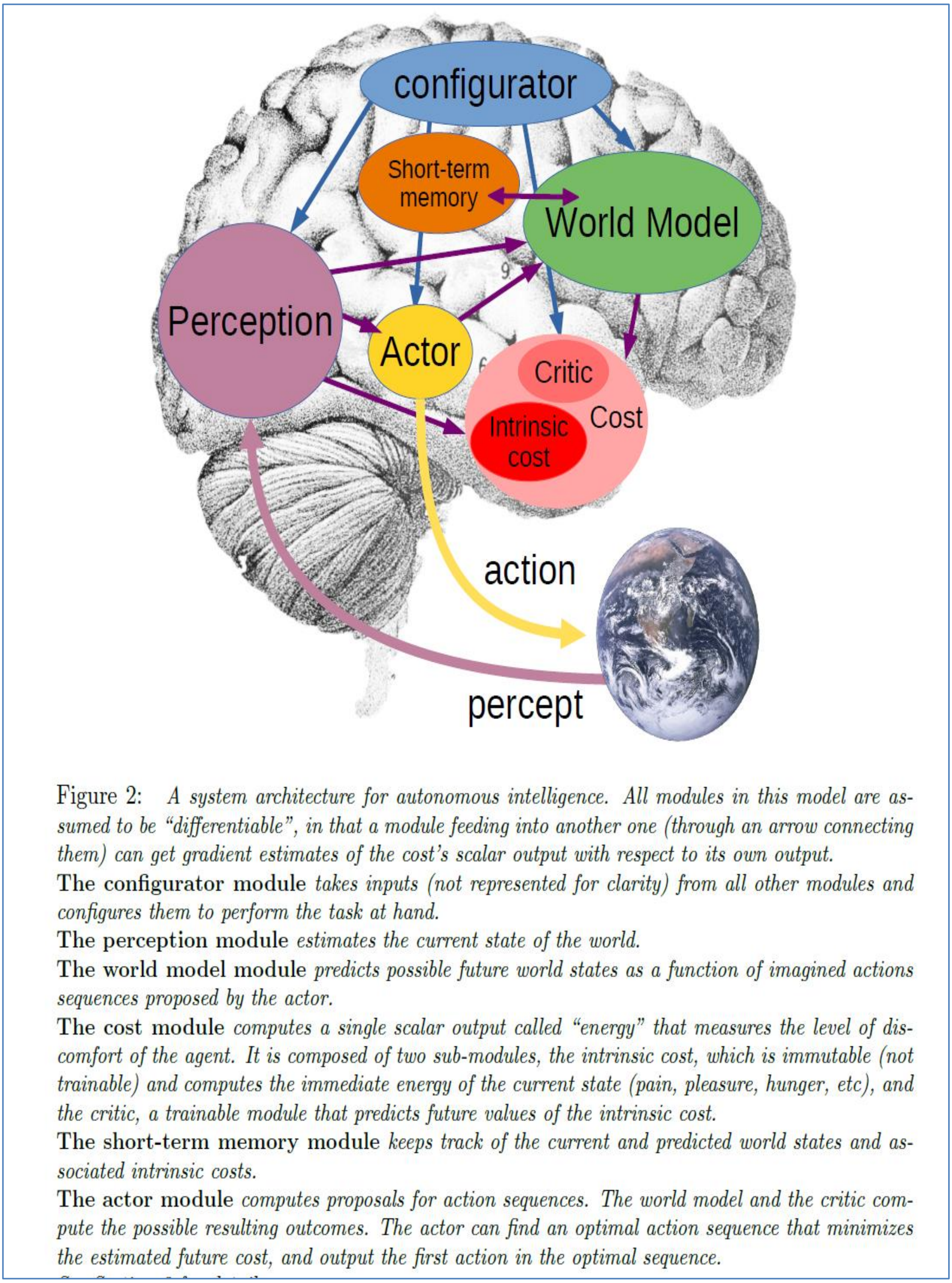


Figure 2: *A system architecture for autonomous intelligence. All modules in this model are assumed to be "differentiable", in that a module feeding into another one (through an arrow connecting them) can get gradient estimates of the cost's scalar output with respect to its own output.*

The configurator module *takes inputs (not represented for clarity) from all other modules and configures them to perform the task at hand.*

The perception module *estimates the current state of the world.*

The world model module *predicts possible future world states as a function of imagined actions sequences proposed by the actor.*

The cost module *computes a single scalar output called "energy" that measures the level of discomfort of the agent. It is composed of two sub-modules, the intrinsic cost, which is immutable (not trainable) and computes the immediate energy of the current state (pain, pleasure, hunger, etc), and the critic, a trainable module that predicts future values of the intrinsic cost.*

The short-term memory module *keeps track of the current and predicted world states and associated intrinsic costs.*

The actor module *computes proposals for action sequences. The world model and the critic compute the possible resulting outcomes. The actor can find an optimal action sequence that minimizes the estimated future cost, and output the first action in the optimal sequence.*

The main contributions of this AMI model are the following:

- An overall cognitive architecture in which all modules are differentiable and many of them are trainable
- JEPA and Hierarchical JEPA: a non-generative architecture for predictive world models that learn a hierarchy of representations
- A non-contrastive self-supervised learning SSL paradigm that produces representations that are simultaneously informative and predictable
- A way to use H-JEPA as the basis of predictive world models for hierarchical planning under uncertainty

Now already – the first real model and real results of this concept:

[Meta AI (2023b)] - Meta AI. I-JEPA: **The first AI model based on Yann LeCun's vision for more human-like AI**, [Assran et al. (2023)] - **Self-Supervised Learning from Images with a Joint-Embedding Predictive**

Last year, Meta's Chief AI Scientist Yann LeCun proposed a new architecture intended to overcome key limitations of even the most advanced AI systems today. His vision is to create machines that can learn internal models of how the world works so that they can learn much more quickly, plan how to accomplish complex tasks, and readily adapt to unfamiliar situations.

We're excited to introduce the first AI model based on a key component of LeCun's vision. This model, the Image Joint Embedding Predictive Architecture (I-JEPA), learns by creating an internal model of the outside world, which compares abstract representations of images (rather than comparing the pixels themselves). I-JEPA delivers strong performance on multiple computer vision tasks, and it's much more computationally efficient than other widely used computer vision models. The representations learned by I-JEPA can also be used for many different applications without needing extensive fine tuning. For example, we train a 632M parameter visual transformer model using 16 A100 GPUs in under 72 hours, and it achieves state-of-the-art performance for low-shot classification on ImageNet, with only 12 labeled examples per class. Other methods typically take two to 10 times more GPU-hours and achieve worse error rates when trained with the same amount of data.

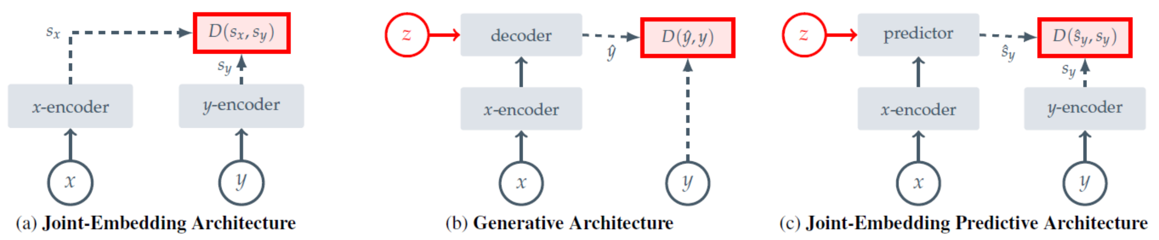


Figure 2. Common architectures for self-supervised learning, in which the system learns to capture the relationships between its inputs. The objective is to assign a high energy (large scaler value) to incompatible inputs, and to assign a low energy (low scaler value) to compatible inputs. **(a)** Joint-Embedding Architectures learn to output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs. **(b)** Generative Architectures learn to directly reconstruct a signal y from a compatible signal x, using a decoder network that is conditioned on additional (possibly latent) variables z to facilitate reconstruction. **(c)** Joint-Embedding Predictive Architectures learn to predict the embeddings of a signal y from a compatible signal x, using a predictor network that is conditioned on additional (possibly latent) variables z to facilitate prediction.

A step closer to human-level intelligence in AI

I-JEPA demonstrates the potential of architectures for learning competitive off-the-shelf image representations without the need for extra knowledge encoded through hand-crafted image transformations. It would be particularly interesting to advance JEPAs to learn more general world-models from richer modalities, e.g., enabling one to make long-range spatial and temporal predictions about future events in a video from a short context, and conditioning these predictions on audio or textual prompts.

We look forward to working to extend the JEPA approach to other domains, like image-text paired data and video data. In the future, JEPA models could have exciting applications for tasks like video understanding. This is an important step towards applying and scaling self-supervised SSL methods for learning a general model of the world.

Key findings for our Project:

A well developed theoretically and methodologically fully functional AI model with "common sense" (general or basic intelligence), while of course not AGI (especially not ASI), but this is a serious step towards it. And now - the first real model and results of this concept already!

And close to our concepts – MAS (Actor+Critic+Configurator) and internal model of the outside world!

What is the role in ASI? - It can be a model for developing the functionality and structures of ASI at different stages of R&D and implementation, and possibly also a subsystem (block) in the ASI itself.

J. Ecosystem of Intelligence from First Principles

In this chapter, we look at the programmatic paper of one of the most influential modern scientists in neurosciences and cognitive science, Karl Friston. He and his team of co-authors propose the concept of a **collective Intelligence** (cyber-physical **ecosystem of intelligent agents** = people + AI) based on the **Active Inference** (adaptive behavior and self-organization based on **the principle of free energy**) with the joint use of a shared generative hyperspatial Bayesian model of the world common to a group of agents and a special communication language.

[Friston et al. (2022)] Karl J. Friston, Maxwell JD Ramstead, Alex B. Kiefer, Alexander Tschantz, Christopher L. Buckley, Mahault Albarracin, Riddhi J. Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton A. R. Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G. Fox, Steven Swanson, Dan Mapes, and Gabriel René. **Designing Ecosystems of Intelligence from First Principles.** arXiv **: 2212.01354v1 [cs.AI] 2** Dec 2022

As in the previous chapters, we first present the structure of the work in the form of a table of contents:

Table of contents

Summary

Active Inference is presented as an approach to AI research and development R&D with the aim of developing **ecosystems of natural and artificial intelligences.**

This approach to General Intelligence (***and hence AGI***) will likely require an understanding of networked or collective intelligence. The zenith of AI could be in the form of **a distributed network of intelligent systems with real-time interaction and composition of emerging forms of intelligence at super-ordinate scales**. **The nodes of this ecosystem can be both people and AI artifacts developed by them.**

Active inference combines the benefits of **First Principles, a physics-based approach to AI with Bayesian formulations, and Bayesian-based machine learning techniques at** the heart of modern AI systems. Active inference explicates **the mechanics of beliefs of agents and groups - Bayesian mechanics** [Ramstead et al. (2022)] – with is uniquely suited to the engineering of intelligent ecosystems, and allows us to describe **the dynamics of spatially connected systems that self-organize at several scale levels (multiscale)**. [Friston] et al. (2015), Friston (2013), Ramstead et al. (2021)].

This encompasses cognition (problem solving through actions and perceptions) and curiosity, as well as creativity fueling the current interest in generative AI [Sequoia Cap. (2022)]. **The design of intelligent systems must begin from the physicality of information and its processing at every scale or level of self-organization.** ***(Stratification from the material and structural stratum - NAE)***

It is necessary to design an AI ecosystem using Active Inference, with a focus on the problem of communication between intelligent agents, with the sharing of forms by the intelligence that arises from these interactions. The Paper highlights also **the importance of shared narratives and goals in the emergence of collective behavior; and how Active Inference helps account for this in terms of sharing (aspects) the same generative model**.

The hypothesis - **to embrace the multi-scale and multi-level aspects of intelligence** - has the potential to be transformable given the assumptions and goals of AI research, development, and design. Technologies based on the described principles can be adapted for the design of emerging intelligence ecosystems covering spatial and cognitive domains (hyperspace networks).

Developing a cyber-physical network of emergent intelligence in the manner described above not only ought to, but for architectural reasons must, be pursued in a way that **positively values and safeguards the individuality of people (as well as potentially non-human persons**).

Proposal for stages of development for active inference as an artificial intelligence technology

AGI and ASI will emerge from the interaction of intelligences networked into a hyper-spatial web or ecosystem of natural and artificial intelligence. Active Inference is proposed as a technology uniquely suited to **the collaborative design of an ecosystem of natural and synthetic sensemaking,** in which humans are integral participants—what we call shared intelligence. The Bayesian mechanics of intelligent systems that follows from active inference led us to define intelligence operationally, as **the accumulation of evidence for an agent's generative model of their sensed world—also known as self-evidencing. This self-evidencing can be implemented using message passing or belief propagation on (factor) graphs or networks.** Active inference is uniquely suited to this task because it leads to a formal account of collective intelligence.

Authors considered that the kinds of communication protocols must be developed to enable (*turn on - initiate)* such an ecosystem of intelligences and argued that such considerations motivate **the development of a generalized, hyper-spatial modeling language and transaction protocol. Establishing such common languages and protocols is a key enabling step towards an ecosystem of naturally occurring and AI. *(Accordingly - and to the initiation of the collective ASI - NAE)***

Stages of development for active inference:

0. **S0: Systemic Intelligence.** This is contemporary state-of-the-art AI; namely, universal function approximation—mapping from input or sensory states to outputs or action states— that optimizes some well-defined value function or cost of (systemic) states.
1. **S1: Sentient Intelligence.** Sentient behavior or Active Inference based on belief updating and propagation (i.e., optimizing beliefs about states as opposed to states per se); where "sentient" means "responsive to sensory impressions."
2. **S2: Sophisticated Intelligence.** Sentient behavior—as defined under S1—in which plans are predicated on the consequences of action for beliefs about states of the world, as opposed to states per se. I.e., a move from "what will happen if I do this?" to "what will I believe or know if I do this?" [Friston et al. (2021), Hesp et al. (2020)].
3. **S3: Sympathetic (or Sapient) Intelligence.** The deployment of sophisticated AI to recognize the nature and dispositions of users and other AI and—in consequence—recognize (and instantiate) attentional and dispositional states of self; namely, a kind of minimal selfhood (which entails generative models equipped with the capacity for Theory of Mind).
4. **S4: Shared (or Super) Intelligence.** (***ASI***) The kind of collective that emerges from the coordination of Sympathetic Intelligence (as defined in S3) and their interaction partners or users—which may include naturally occurring intelligence such as ourselves, but also other sapient artifacts. We believe that the approach that we have outlined here is the most likely route toward this kind of hypothetical, planetary-scale, distributed superintelligence [Frank et al. (2022)]. ***(Here comes SkyNet!!! – NAE)***

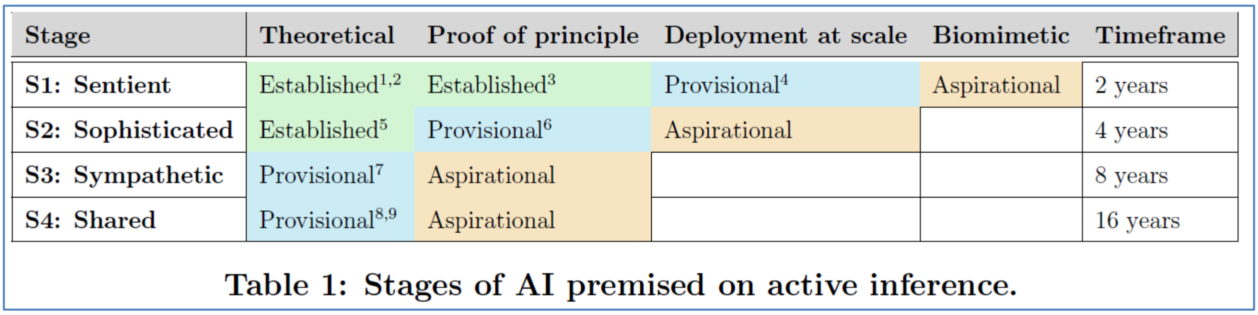
Implementation

A. **Theoretical**. The basis of belief updating (i.e., inference and learning) is underwritten by a formal calculus (e.g., Bayesian mechanics), with clear links to the physics of self-organization of open systems far from equilibrium.

B. **Proof of principle.** Software instances of the formal (mathematical) scheme, usually on a classical (i.e., von Neumann) architecture.

C. **Deployment at scale.** Scaled and efficient application of the theoretical principles (i.e., methods) in a real-world setting (e.g., edge-computing, robotics, variational message passing on the web, etc.)

D. **Biomimetic hardware.** Implementations that elude the von Neumann bottleneck, on biomimetic or neuromorphic architectures. E.g., photonics, soft robotics, and belief propagation: i.e., message passing of the sufficient statistics of (Bayesian) beliefs.



Stage	Theoretical	Proof of principle	Deployment at scale	Biomimetic	Timeframe
S1: Sentient	Established[1,2]	Established[3]	Provisional[4]	Aspirational	2 years
S2: Sophisticated	Established[5]	Provisional[6]	Aspirational		4 years
S3: Sympathetic	Provisional[7]	Aspirational			8 years
S4: Shared	Provisional[8,9]	Aspirational			16 years

Table 1: Stages of AI premised on active inference.

1. [Friston (2019)]
2. [Ramstead et al. (2022)]
3. [Parr et al. (2022)]
4. [Mazzaglia et al. (2022)]
5. [Da Costa et al. (2020)]
6. [Friston et al. (2017)]
7. [Friston et al. (2020)]
8. [Friston et al. (2015)]
9. [Albarracin et al. (2022)]

[Isomura et al. (2023)] – **New empirical confirmation of this concept.**

Empirical applications of the free-energy principle are not straightforward because they entail a commitment to a particular process theory, especially at the cellular and synaptic levels. Using a recently established reverse engineering technique, we confirm the quantitative predictions of the free-energy principle using in vitro networks of rat cortical neurons that perform causal inference. Upon receiving electrical stimuli—generated by mixing two hidden sources—neurons self-organised to selectively encode the two sources. Pharmacological up- and downregulation of network excitability disrupted the ensuing inference, consistent with changes in prior beliefs about hidden sources. As predicted, changes in effective synaptic connectivity reduced variational free energy, where the connection strengths encoded parameters of the generative model. In short, we show that variational free energy minimisation can quantitatively predict the self-organisation of neuronal networks, in terms of their responses and plasticity. These results demonstrate the applicability of the free-energy principle to in vitro neural networks and establish its predictive validity in this setting.

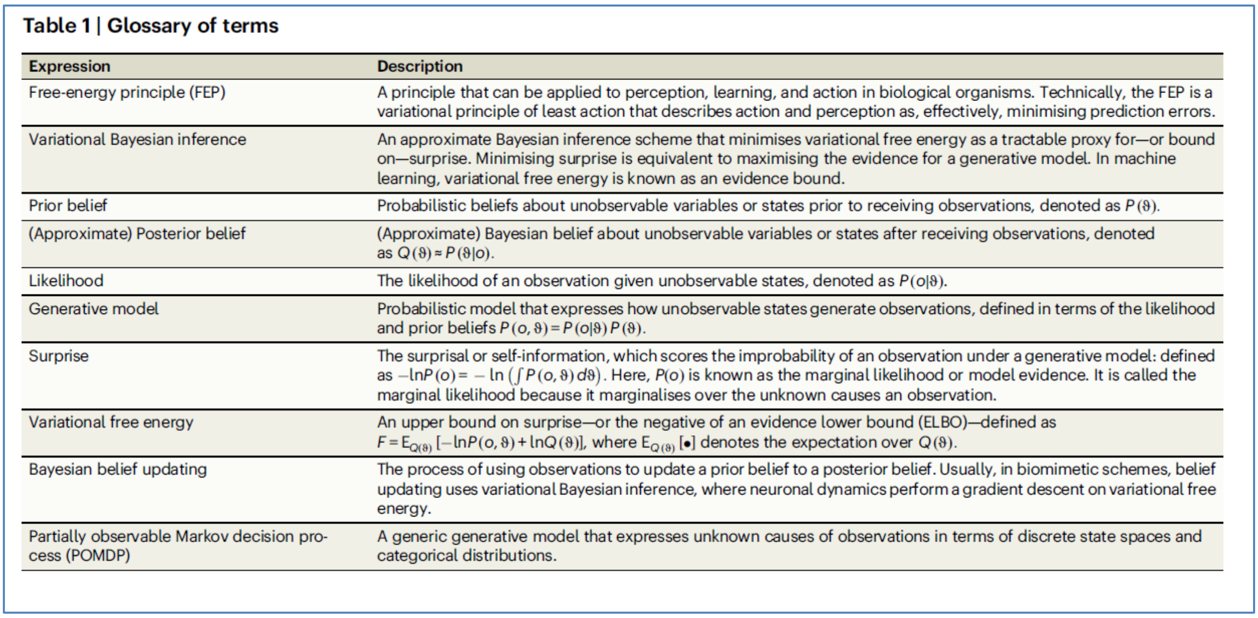
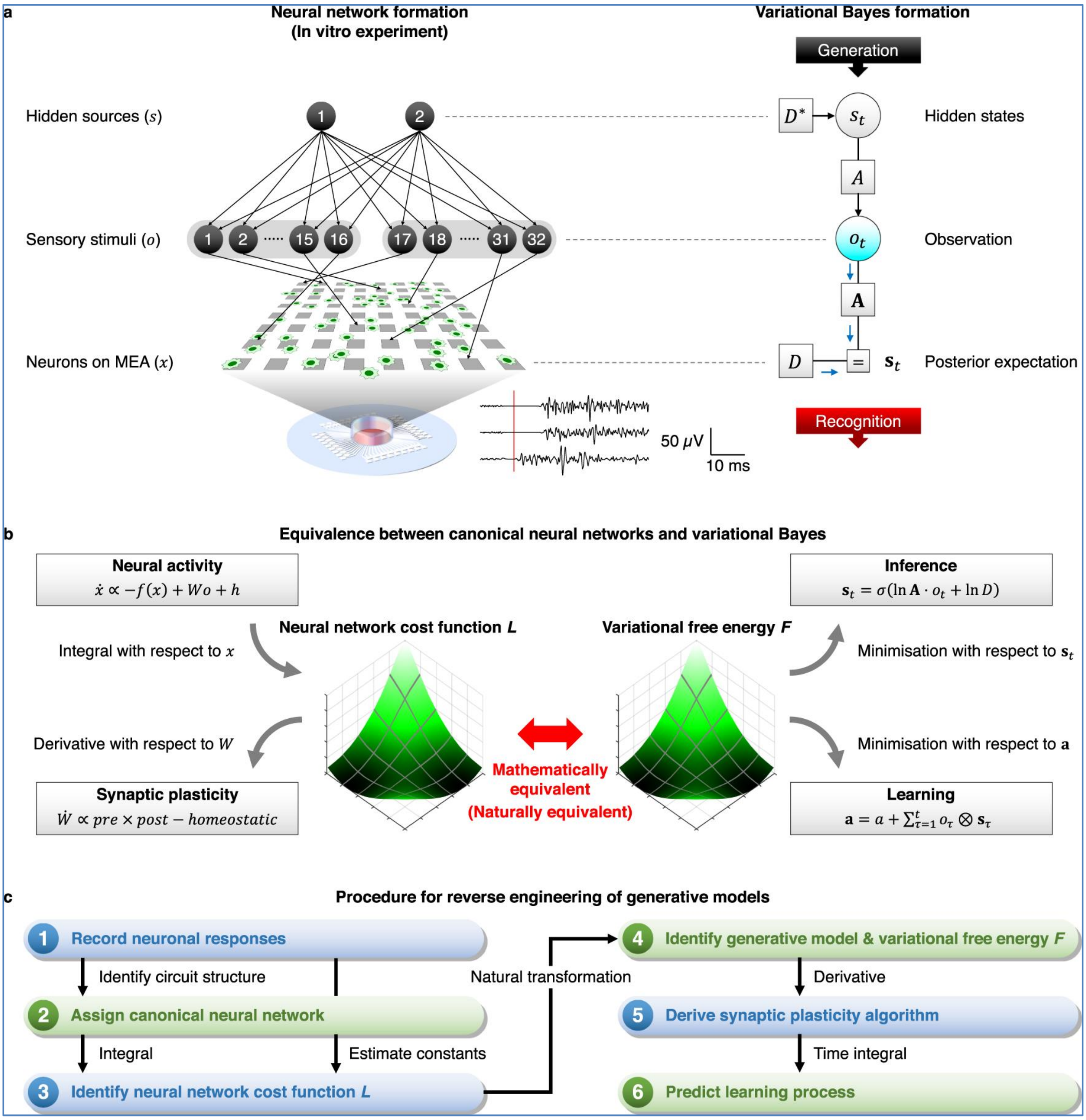


Table 1 | Glossary of terms

Expression	Description
Free-energy principle (FEP)	A principle that can be applied to perception, learning, and action in biological organisms. Technically, the FEP is a variational principle of least action that describes action and perception as, effectively, minimising prediction errors.
Variational Bayesian inference	An approximate Bayesian inference scheme that minimises variational free energy as a tractable proxy for—or bound on—surprise. Minimising surprise is equivalent to maximising the evidence for a generative model. In machine learning, variational free energy is known as an evidence bound.
Prior belief	Probabilistic beliefs about unobservable variables or states prior to receiving observations, denoted as $P(\vartheta)$.
(Approximate) Posterior belief	(Approximate) Bayesian belief about unobservable variables or states after receiving observations, denoted as $Q(\vartheta) \approx P(\vartheta \mid o)$.
Likelihood	The likelihood of an observation given unobservable states, denoted as $P(o \mid \vartheta)$.
Generative model	Probabilistic model that expresses how unobservable states generate observations, defined in terms of the likelihood and prior beliefs $P(o, \vartheta) = P(o \mid \vartheta)\, P(\vartheta)$.
Surprise	The surprisal or self-information, which scores the improbability of an observation under a generative model: defined as $-\ln P(o) = -\ln\left(\int P(o, \vartheta)\, d\vartheta\right)$. Here, $P(o)$ is known as the marginal likelihood or model evidence. It is called the marginal likelihood because it marginalises over the unknown causes an observation.
Variational free energy	An upper bound on surprise—or the negative of an evidence lower bound (ELBO)—defined as $F = E_{Q(\vartheta)}[-\ln P(o, \vartheta) + \ln Q(\vartheta)]$, where $E_{Q(\vartheta)}[\bullet]$ denotes the expectation over $Q(\vartheta)$.
Bayesian belief updating	The process of using observations to update a prior belief to a posterior belief. Usually, in biomimetic schemes, belief updating uses variational Bayesian inference, where neuronal dynamics perform a gradient descent on variational free energy.
Partially observable Markov decision process (POMDP)	A generic generative model that expresses unknown causes of observations in terms of discrete state spaces and categorical distributions.

Fig. 1: Reverse engineering of the generative model from empirical data.


a
Neural network formation
(In vitro experiment)
Variational Bayes formation
Generation
Hidden sources (s)
1
2
D^*
s_t
Hidden states
A
Sensory stimuli (o)
1 2 ····· 15 16 17 18 ····· 31 32
o_t
Observation
$\mathbf{A}$
Neurons on MEA (x)
D
=
$\mathbf{s}_t$
Posterior expectation
Recognition
50 μV
10 ms
b
Equivalence between canonical neural networks and variational Bayes
Neural activity
$\dot{x} \propto -f(x) + Wo + h$
Inference
$\mathbf{s}_t = \sigma(\ln \mathbf{A} \cdot o_t + \ln D)$
Integral with respect to x
Neural network cost function L
Variational free energy F
Minimisation with respect to $\mathbf{s}_t$
Derivative with respect to W
Minimisation with respect to $\mathbf{a}$
Mathematically equivalent
(Naturally equivalent)
Synaptic plasticity
$\dot{W} \propto pre \times post - homeostatic$
Learning
$\mathbf{a} = a + \sum_{\tau=1}^{t} o_\tau \otimes \mathbf{s}_\tau$
c
Procedure for reverse engineering of generative models
1 Record neuronal responses
Identify circuit structure
2 Assign canonical neural network
Integral
Estimate constants
3 Identify neural network cost function L
Natural transformation
4 Identify generative model & variational free energy F
Derivative
5 Derive synaptic plasticity algorithm
Time integral
6 Predict learning process

Key findings and conclusions for our Project – Fristons's concept is close to our!

- Stratification of AI systems, starting with material and structural stratum
- Cybernetic control models CSs in AI systems
- Upgradable models of the world and AI itself
- Self-organization of the ASI system in the environment created for this - an ecosystem
- Semiotics as the basis of communications in AI systems and the ASI ecosystem
- Using quantum computing for belief updating
- Collective ASI - a network/system of agents, including people and AI
- The highest level of ASI Ethics

K. Large Language Models. GPT-4 and others

At the beginning of 2023, Large Language Models (LLMs) were defined (designated) as the most advanced and promising BMs (see Appendix H). The largest and frontier models from industry leaders:

- **PalM2** (Google, USA) – [Google (2023), Chowdhery et al. (2022), Tay et al. (2023), Hoffmann et al. (2022), Lee et al. (2021)]
- **Claude-2** (Anthropic, USA) – [Anthropic 2022), (2023a), (2023b)]
- **Llama 2** (Meta, USA) – [Meta AI (2023a)]
- **Aquila** (BAAI, China) – [BAAI (2023)]
- **ERNIE Bot** (Baidu, China) – [Baidu (2023)]
- **Grok** (xAI, USA) – [xAI (2023)] - and finally from Elon Musk!
- **Gemini** (Google DeepMind, USA) - [Google DeepMind (2023c), (2023d), Pichai & Hassabis (2023)]

Let's take a closer look at the papers on the most famous and successful **GPT-4** model (OpenAI, USA):

GPT-4 reports

[OpenAI (2023a)] - **GPT-4 technical report.**

GPT-4 is a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior.

Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation.

GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cuts off in September 2021, and does not learn from its experience. It can sometimes make simple reasoning errors which do not seem to comport with competence across so many domains, or be overly gullible in accepting obviously false statements from a user. It can fail at hard problems the same way humans do, such as introducing security vulnerabilities into code it produces.

GPT-4 can also be confidently wrong in its predictions, not taking care to double-check work when it’s likely to make a mistake. Interestingly, the pre-trained model is highly calibrated (its predicted confidence in an answer generally matches the probability of being correct).

[OpenAI (2023b)] - **GPT-4 System Card.**

Large language models, also known as LLMs, have become an increasingly prevalent part of our day-to-day lives, with their use extending to a wide range of domains including web browsing, voice assistants, and coding assistance tools.

GPT models are often trained in two stages. First, they are trained, using a large dataset of text from the Internet, to predict the next word. The models are then fine-tuned with additional data, using an algorithm called reinforcement learning from human feedback (RLHF), to produce outputs that are preferred by human labelers.

Some of the specific risks we explored are:

- Hallucinations
- Harmful content
- Harms of representation, allocation, and quality of service
- Disinformation and influence operations
- Proliferation of conventional and unconventional weapons
- Privacy
- Cybersecurity
- Potential for risky emergent behaviors
- Interactions with Other Systems
- Economic impacts
- Acceleration
- Overreliance

[Bubeck et al. (2023)] - **Sparks of Articial General Intelligence: Early experiments with GPT-4**

Articial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data.

Conclusions

GPT-4 attains a form of general intelligence, indeed showing sparks of artificial general intelligence. This is demonstrated by its core mental capabilities (such as reasoning, creativity, and deduction), its range of topics on which it has gained expertise (such as literature, medicine, and coding), and the variety of tasks it is able to perform (e.g., playing games, using tools, explaining itself etc.). A lot remains to be done to create a system that could qualify as a complete AGI. We conclude this paper by discussing several immediate next steps, regarding defining AGI itself, building some of missing components in LLMs for AGI, as well as gaining better understanding into the origin of the intelligence displayed by the recent LLMs.

On the path to more general artificial intelligence

- Confidence calibration
- Long-term memory LTM
- Continual learning

- Personalization
- Planning and conceptual leaps
- Transparency, interpretability and consistency
- Cognitive fallacies and irrationality
- Challenges with sensitivity to inputs

Potential extensions to next word prediction include the following:

- External calls by the model to components and tools such as a calculator, a database search or code execution
- A richer, more complex “slow-thinking" deeper mechanism that oversees the “fast-thinking" mechanism of next word prediction
- Integration of long-term memory LTM as an inherent part of the architecture, perhaps in the sense that both the input and output of the model will include, in addition to the tokens representing the text, a vector which represents the context
- Going beyond single-word prediction: Replacing the sequence of tokens by a hierarchical structure, where higher-level parts of the text such as sentences, paragraphs or ideas are represented in the embedding and where the content is generated in a top-down manner.

[Hoffman & GPT-4 (2023)] - **Impromptu. Amplifying Our Humanity Through AI.**

Large Language Models like GPT-4, can elevate humanity across key areas like education, business, justice, journalism, social media and creativity. Reid Hoffman explores the current state of AI and its potential to amplify our humanity and offers a unique perspective on the impact of AI on our lives.

LLMs Research & Development R&D

[Shanahan et al. (2023)] **– Role-Play with LLMs**

As dialogue agents become increasingly humanlike in their performance, it is imperative that we have to develop effective ways to describe their behavior in high-level terms without falling into the trap of anthropomorphism. This paper foregrounds the concept of role-play. Casting dialogue agent behavior in terms of role-play allows us to draw on familiar folk psychological terms, without ascribing human characteristics to LLMs they in fact lack.

[OpenAI (2023c)] - **Improving mathematical reasoning with process supervision.**

We've trained a model to achieve a new state-of-the-art in mathematical problem solving by rewarding each correct step of reasoning (“process supervision”) instead of simply rewarding the correct final answer (“outcome supervision”). In addition to boosting performance relative to outcome supervision, process supervision also has an important alignment benefit: it directly trains the model to produce a chain-of-thought that is endorsed by humans.

[Lightman et al. (2023)] - **Let’s Verify Step by Step**

In recent years, large language models LLMs have greatly improved in their ability to perform complex multi-step reasoning. However, even stateof-the-art models still regularly produce logical mistakes. To train more reliable models, we can turn either to outcome supervision, which provides feedback for a final result, or process supervision, which provides feedback for each intermediate reasoning step. Given the importance of training reliable models, and given the high cost of human feedback, it is important to carefully compare the both methods. Recent work has already begun this comparison, but many questions still remain. We conduct our own investigation, finding that process supervision significantly outperforms outcome supervision for training models to solve problems from the challenging MATH dataset. Our process-supervised model solves 78% of problems from a representative subset of the MATH test set. Additionally, we show that active learning significantly improves the efficacy of process supervision. To support related research, we also release PRM800K, the complete dataset of 800,000 step-level human feedback labels used to train our best reward model.

[Hu & Clune (2023)] - **Thought Cloning: Learning to Think while Acting by Imitating Human Thinking.**

Language is often considered a key aspect of human thinking, providing us with exceptional abilities to generalize, explore, plan, replan, and adapt to new situations. However, Reinforcement Learning (RL) agents are far from human-level performance in any of these abilities. We hypothesize one reason for such cognitive deficiencies is that they lack the benefits of thinking in language and that we can improve AI agents by training them to think like humans do. We introduce a novel Imitation Learning framework, Thought Cloning, where the idea is to not just clone the behaviors of human demonstrators, but also the thoughts humans have as they perform these behaviors. While we expect Thought Cloning to truly shine at scale on internet-sized datasets of humans thinking out loud while acting (e.g. online videos with transcripts), here we conduct experiments in a domain where the thinking and action data are synthetically generated.

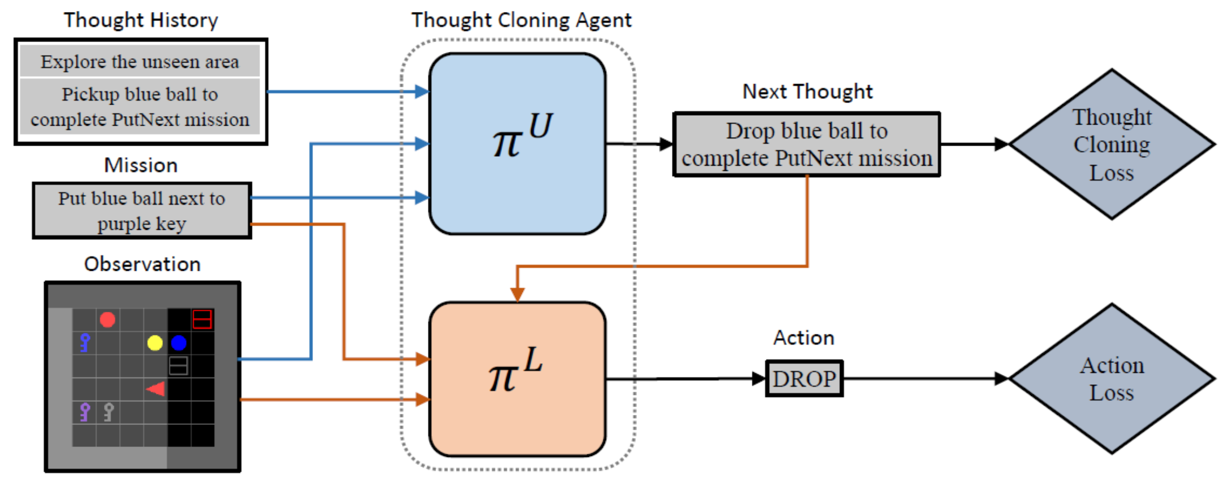


Figure 1: Overall framework for Thought Cloning (TC). The TC agent has two components: the Upper-Level and Lower-level Components. At each timestep, the TC agent receives an observation, a mission, and a history of thoughts as inputs. The Upper-Level Component generates thoughts, and the Lower-Level Component generates actions conditioned on these thoughts. Generated thoughts and actions are compared to the ground truth from the demonstration dataset to calculate the loss.

Results reveal that Thought Cloning learns much faster than Behavioral Cloning and its performance advantage grows the further out of distribution test tasks are, highlighting its ability to better handle novel situations. Thought Cloning also provides important benefits for AI Safety and Interpretability, and makes it easier to debug and improve AI. Because we can observe the agent's thoughts, we can (1) more easily diagnose why things are going wrong, making it easier to fix the problem, (2) steer the agent by correcting its thinking, or (3) prevent it from doing unsafe things it plans to do. By training agents how to think as well as behave, Thought Cloning creates safer, more powerful agents.

[OpenAI (2023d)] - **Introducing Superalignment**

We need scientific and technical breakthroughs to steer and control AI systems much smarter than us. To solve this problem within four years, we're starting a new team, co-led by Ilya Sutskever and Jan Leike, and dedicating 20% of the compute we've secured to date to this effort.

Our goal is to solve the core technical challenges of superintelligence alignment in four years.

[Li et al. (2022)] **- A systematic investigation of commonsense knowledge in large language models LLMs.**

[Wang et al. (2023)] - **Describe, Explain, Plan and Select: Interactive planning with large language models LLMs enables open-world multi-task agents.**

[Lin et al. (2023)] - **Text2Motion: From natural language instructions to feasible plans.**

[Webb et al. (2023)] - **Emergent Analogical Reasoning in Large Language Models**

The recent advent of large language models LLMs has reinvigorated debate over whether human cognitive capacities might emerge in such generic models given sufficient training data. Of particular interest is the ability of these models to reason about novel problems zero-shot, without any direct training. In human cognition, this capacity is closely tied to an ability to reason by analogy. Here, we performed a direct comparison between human reasoners and a large language model (the text-davinci-003 variant of GPT-3) on a range of analogical tasks, including a non-visual matrix reasoning task based on the rule structure of Raven's Standard Progressive Matrices. We found that GPT-3 displayed a surprisingly strong capacity for abstract pattern induction, matching or even surpassing human capabilities in most settings; preliminary tests of GPT-4 indicated even better performance. Our results indicate that large language models such as GPT-3 have acquired an emergent ability to find zero-shot solutions to a broad range of analogy problems.

[Gurnee & Tegmark (2023)] findings from MIT: **LLMs** (Llama-2 family) **represent Space and Time!**

[Google DeepMind (2023a)] - **PROMPTBREEDER, a general-purpose self-referential selfimprovement mechanism that evolves and adapts prompts for a given domain.**

Driven by an LLM, Promptbreeder mutates a population of task-prompts, evaluates them for fitness on a training set, and repeats this process over multiple generations to evolve task-prompts. Crucially, the mutation of these task-prompts is governed by mutation-prompts that the LLM generates and improves throughout evolution in a self-referential way. That is, Promptbreeder is not just improving task-prompts, but it is also improving the mutation-prompts that improve these task-prompts. Promptbreeder outperforms state-of-the-art prompt strategies such as Chain-of-Thought and Plan-and-Solve Prompting on commonly used arithmetic and commonsense reasoning benchmarks. Furthermore, Promptbreeder is able to evolve intricate task-prompts for the challenging problem of hate speech classification.

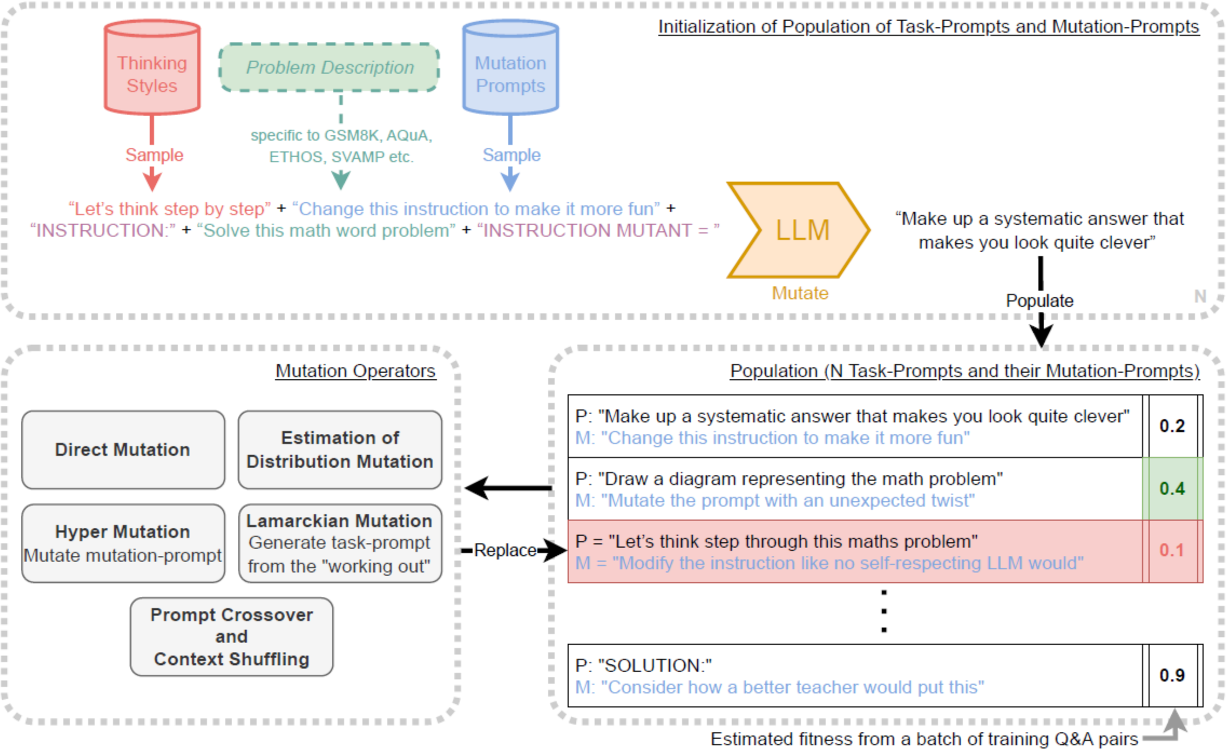

Initialization of Population of Task-Prompts and Mutation-Prompts
Thinking Styles
Problem Description
Mutation Prompts
Sample
specific to GSM8K, AQuA, ETHOS, SVAMP etc.
Sample
"Let's think step by step" + "Change this instruction to make it more fun" + "INSTRUCTION:" + "Solve this math word problem" + "INSTRUCTION MUTANT = "
LLM
Mutate
"Make up a systematic answer that makes you look quite clever"
Populate
N
Mutation Operators
Direct Mutation
Estimation of Distribution Mutation
Hyper Mutation
Mutate mutation-prompt
Lamarckian Mutation
Generate task-prompt from the "working out"
Prompt Crossover and Context Shuffling
Replace
Population (N Task-Prompts and their Mutation-Prompts)
P: "Make up a systematic answer that makes you look quite clever"
M: "Change this instruction to make it more fun"
0.2
P: "Draw a diagram representing the math problem"
M: "Mutate the prompt with an unexpected twist"
0.4
P = "Let's think step through this maths problem"
M = "Modify the instruction like no self-respecting LLM would"
0.1
P: "SOLUTION:"
M: "Consider how a better teacher would put this"
0.9
Estimated fitness from a batch of training Q&A pairs

Figure 1: Overview of Promptbreeder. Given a problem description and an initial set of general "thinking-styles" and mutation-prompts, Promptbreeder generates a population of units of evolution, each unit consisting of typically two task-prompts and a mutation-prompt. We then run a standard binary tournament genetic algorithm (Harvey, 2011). To determine the fitness of a task-prompt we evaluate its performance on a random batch of training data. Over multiple generations, Promptbreeder subsequently mutates task-prompts as well as mutation-prompts using five different classes of mutation operators. The former leads to increasingly domain-adaptive task-prompts whereas the latter evolves increasingly useful mutation-prompts in a self-referential way.

[Perez (2023)] **– New promising methods and tools for LLMs development**

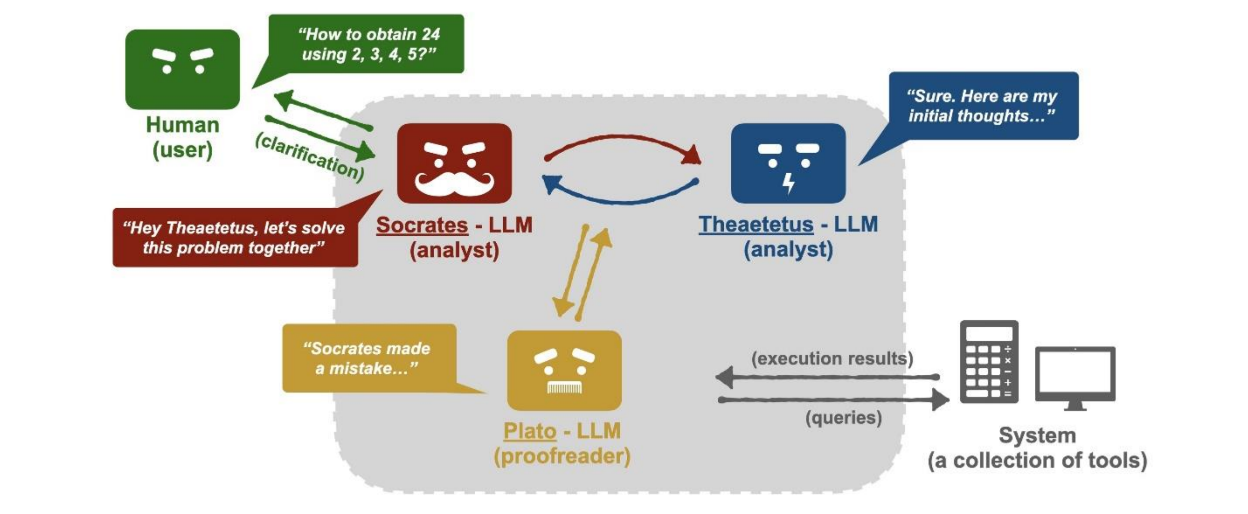
Retrieval-augmented generation (RAG) powers modern chatbots to **handle real-world open-domain conversations** and has become popular for knowledge-intensive NLP tasks.

Thread of Thought (ThoT), an elegant prompting strategy that structures LLMs to **methodically analyze chaotic retrieved contexts.**

System 2 Attention (S2A) - S2A uses the generative capabilities of LLMs, prompting them **to regenerate only relevant context by removing distractions**. It's a way to embed attention control right into the prompt with a reasoning-based rewrite, rather than relying solely on output treatments. . More detailed see paper from Meta researches [Weston & Sukhbaatar (2023)]

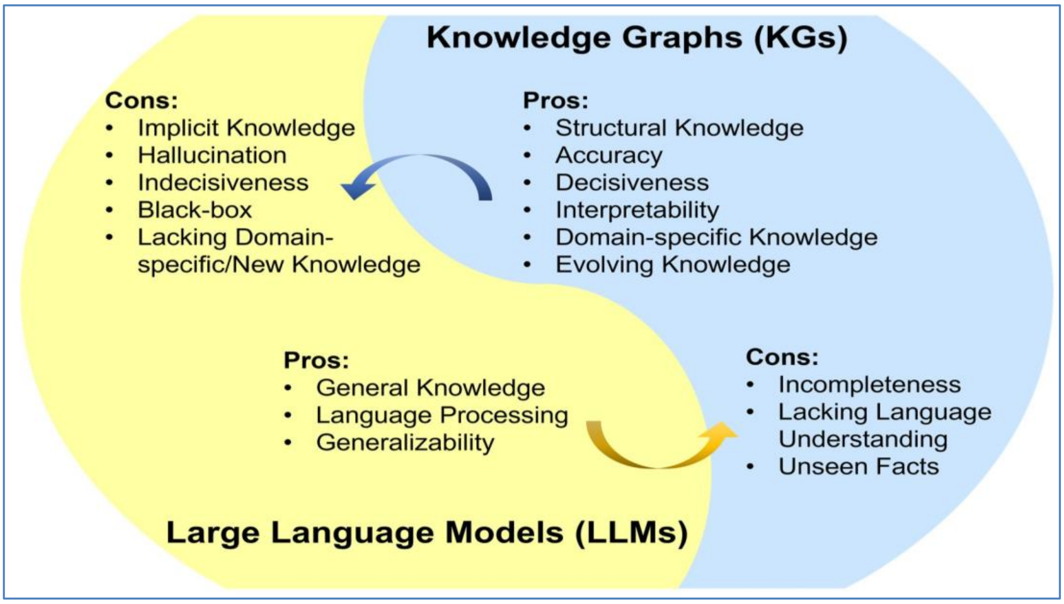
Learning from Mistakes (LEMA) training - create LLMs that augment their reasoning skills **by identifying flaws in their logic, explaining why they were wrong, and correcting their own mistakes (*feedback*!)**. It gains a "consciousness" about the principles of mathematical reasoning.

SocraticAI simulates fluid human discussion through three distinct AI agents - Socrates, Theaetetus, and Plato. (*MAS!)* SocraticAI allows AI to truly learn through dialogue - questioning, explaining, and building upon new insights as they emerge.

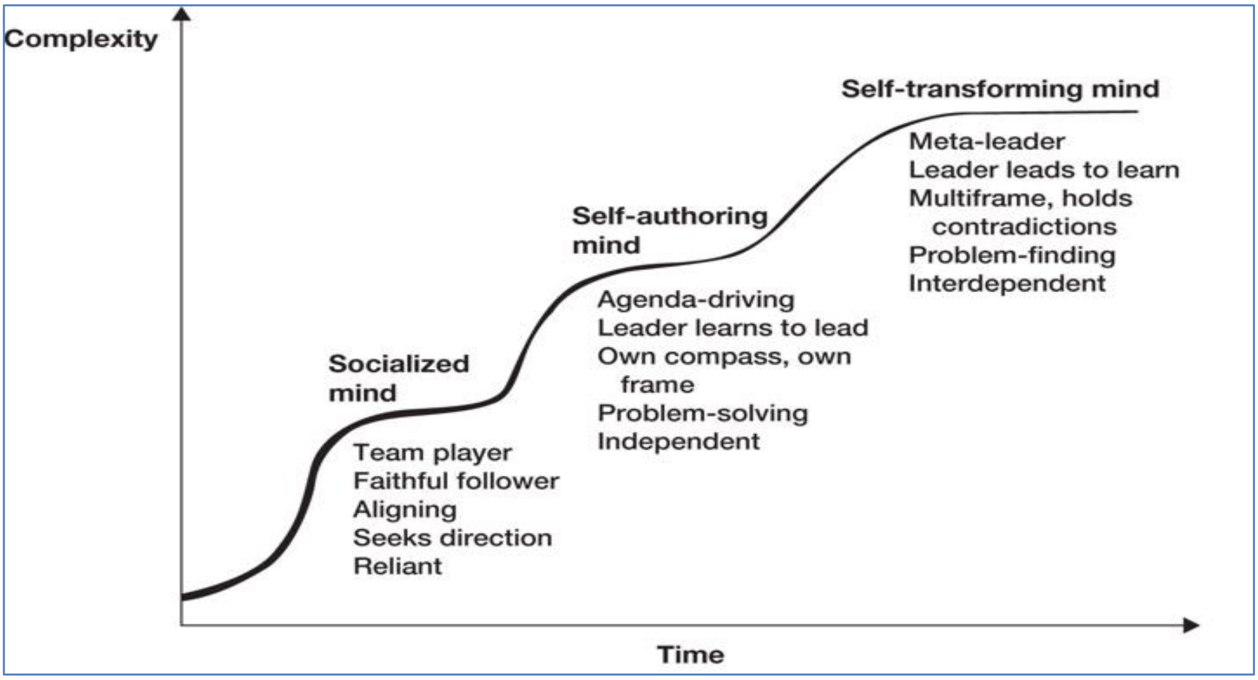

"How to obtain 24 using 2, 3, 4, 5?"
Human (user)
(clarification)
"Sure. Here are my initial thoughts…"
"Hey Theaetetus, let's solve this problem together"
Socrates - LLM (analyst)
Theaetetus - LLM (analyst)
"Socrates made a mistake…"
Plato - LLM (proofreader)
(execution results)
(queries)
System (a collection of tools)

Socratic AI: a framework for collaborative problem-solving with LLMs

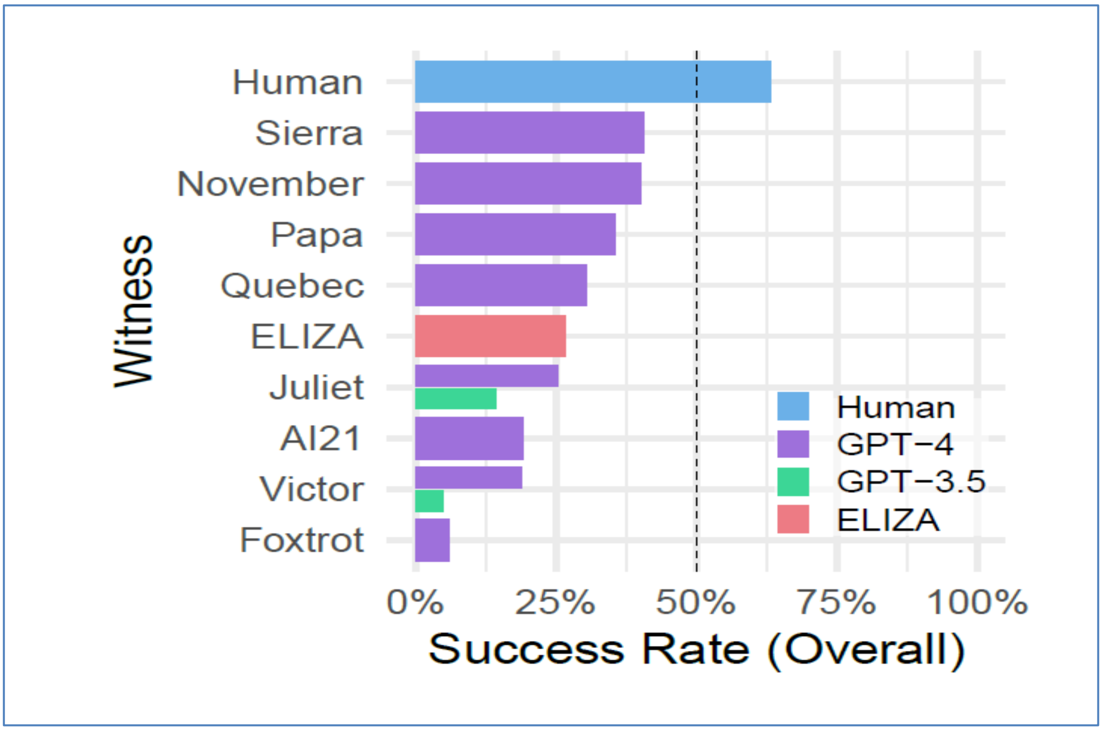
Large language models (LLMs) and knowledge graphs (KGs) are complementary technologies that balance each other's strengths and weaknesses when combined



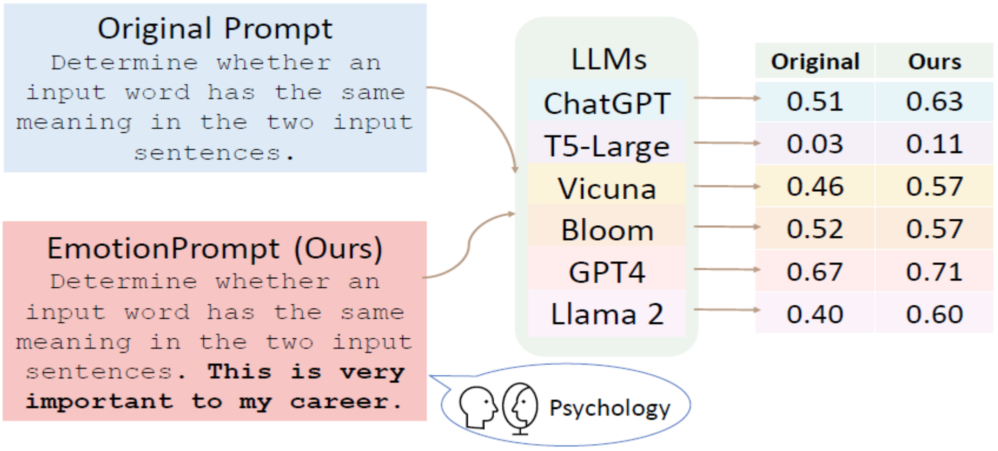
FINALLY form Carlos Perez: “This is actually a much bigger a deal because GPT can now retrieve information that is *not* in its knowledge on the fly! It implies a first step towards an LLM that is not unencumbered by its original training set! It's a first step in a self-authoring mind.”



[Jones & Bergen (2023)] - **Does GPT-4 Pass the Turing Test?** – **YES, LLM PASSED**! With best result 41%, it matches of criteria formulated by Alan Turing itself (>30%). But still less then human (63%)…



[Li et al. (2023)] - **Large Language Models Understand and Can Be Enhanced by Emotional Stimuli – YES, LLM HAS (some) Emotional intelligence!**


Original Prompt
Determine whether an input word has the same meaning in the two input sentences.
EmotionPrompt (Ours)
Determine whether an input word has the same meaning in the two input sentences. This is very important to my career.
Psychology

LLMs	Original	Ours
ChatGPT	0.51	0.63
T5-Large	0.03	0.11
Vicuna	0.46	0.57
Bloom	0.52	0.57
GPT4	0.67	0.71
Llama 2	0.40	0.60

Figure 1: An overview of our research from generating to evaluating EmotionPrompt.

Summary, key findings and conclusions for our Project

- **Architecture - a neural network-transformer**, that is, capable of adapting to any new tasks
- **Generative** - capable and intended to generate new content – text (now – mulimodal!)
- **Pre-training** - pre-trained on huge amounts of raw data (see chapter 50. Data for BMs) and (almost) do not require additional special training
- **Universal** (multipurpose) in use due to pre-training
- **Multimodal** - not only text requests, but also any modalities can be received as input
- **Multilingual** - use any language (level depends on data availability)
- Able to use a sufficiently **large amount of context** on the input
- **Interfaces** - natural language text chat and Application Programming Interface API (that is, the ability to interact with other programs and applications)
- **Multi-user** - work simultaneously with many users
- **The closest to AGI** - emergence, reasoning, some "common sense" etc.

The problems and limitations of LLMs are correctable during refinement and it is clear how - they are already working on it and this work is ongoing:

- **Scalability and non-linear development**
- **Long Term Memory LTM**
- **Knowledge Graphs KGs**
- **Feedback control algorithms**
- **Step by step control and checking**
- **Collaboration with external applications via API**
- **Online access to the Internet and other data**
- **Training based on current work - that is, on own (collected) self experience**
- **MAS with separation of functions and mutual control**

L. Consciousness in Artificial Intelligence

This topic is based on significant paper – [**Butlin et al. (2023)]** Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, Rufin VanRullen. **Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.** arXiv:2308.08708v3 [cs.AI] 22 Aug 2023

Abstract

Whether current or near-term AI systems could be conscious is a topic of scientific interest and increasing public concern. This report argues for, and exemplifies, a rigorous and empirically grounded approach to AI consciousness: assessing existing AI systems in detail, in light of our best-supported neuroscientific theories of consciousness. We survey several prominent scientific theories of consciousness, including recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory. From these theories we derive "indicator properties" of consciousness, elucidated in computational terms that allow us to assess AI systems for these properties. We use these indicator properties to assess several recent AI systems, and we discuss how future systems might implement them. Our analysis suggests that no current AI systems are conscious, but also suggests that there are no obvious technical barriers to building AI systems which satisfy these indicators.

Contents

<u>Several key definitions from this useful Glossary</u>

- **access consciousness** - “Functional” concept contrasted with **phenomenal consciousness**; a state is access conscious if its content is directly available to its subject to perform a wide range of cognitive tasks such as report, reasoning, and rational action
- **computational functionalism** - The thesis that implementing computations of a certain kind is necessary and sufficient for consciousness
- **first-order representations** - Representations that are about the non-representational world, in contrast with **higher-order representations**; paradigm cases include the visual representation of an external object like an apple
- **higher-order representations** - Representations that are about other representations (e.g. a representation that another representation is reliable)
- **metacognition** - Cognition about one’s own cognitive processes, for example about their reliability or accuracy
- **phenomenal consciousness** - Consciousness as we understand it in this report
- **theory-heavy approach** - Method for determining which systems are conscious based on scientific theories of consciousness

Methods and Assumptions

In using the term "phenomenal consciousness", we mean to distinguish our topic from "access consciousness", following [Block (1995), (2002)]. Block writes that "a state is (access conscious) if it is broadcast for free use in reasoning and for direct 'rational' control of action (including reporting)"

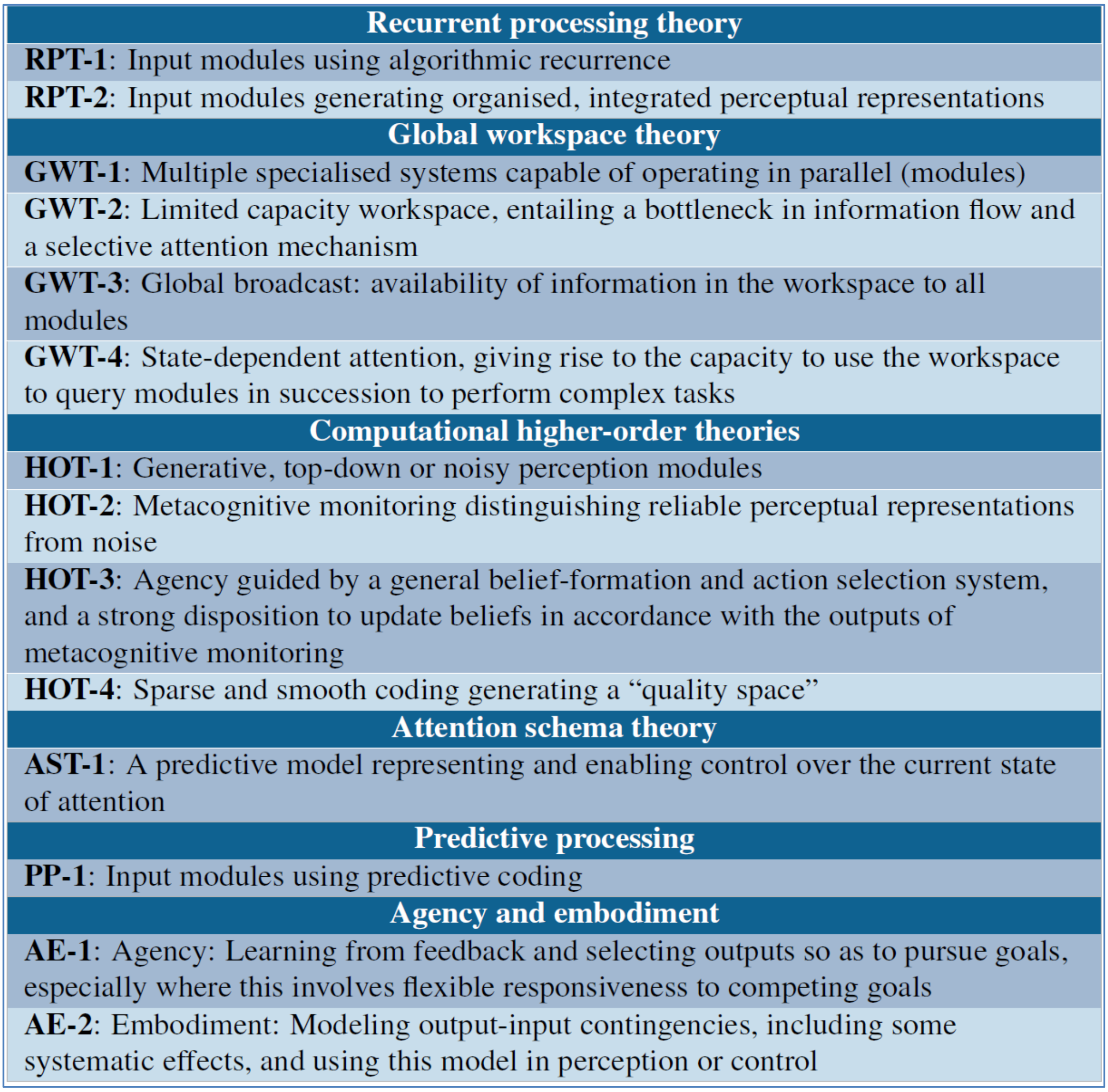
- **Computational functionalism**: Implementing computations of a certain kind is necessary and sufficient for consciousness, so it is possible in principle for non-organic artificial systems to be conscious [Block (1996), (2023)].
- **Scientific theories**: Neuroscientific research has made progress in characterising functions that are associated with, and may be necessary or sufficient for, consciousness; these are described by scientific theories of consciousness [Seth & Bayne (2022), Yaron et al. (2022)].
- **Theory-heavy approach:** A particularly promising method for investigating whether AI systems are likely to be conscious is assessing whether they meet functional or architectural conditions drawn from scientific theories, as opposed to looking for theory-neutral behavioural signatures [Birch (2022b)].

Two further points about our methods and assumptions are worth noting before we go on. The first is that, for convenience, we will generally write as though whether **a system is conscious is an all-or-nothing matter, and there is always a determinate fact about this** (although in many cases this fact may be difficult to learn). However, we are open to the possibility that this may not be the case: that it may be possible for a system to be partly (and in multiple dimensions) conscious [Birch et al. (2020)], conscious to some degree [Lee (2022), Shulman & Bostrom (2021)], or neither determinately conscious nor determinately non-conscious [Birch (2022a), Simon (2017) - Schwitzgebel forthcoming]

Theories and Concepts

- Recurrent Processing Theory RPT – [Lamme (2006), (2010), (2020)]
- Global Workspace Theory GWT – [Dehaene et al. (1998), (2003), Dehaene & Naccache (2001), Dehaene & Changeux (2011), Dehaene (2014), Mashour et al. (2020)]
- Higher-Order Theories HOT - [Brown et al. (2019), Fleming (2020), Lau (2019), (2022) - Michel forthcoming]
- Attention Schema Theory AST – [Webb & Graziano (2015), Graziano (2019), Liu et al. (2023)]
- Predictive Processing PP – [Deane (2021), Hohwy (2022), Nave et al. (2022), Friston (2010), Whyte (2019), Fleming (2020)]
- Agency and Embodiment AE – [Dolan & Dayan (2013), Russell & Norvig (2021), Godfrey-Smith (2016), (2019), Hohwy (2022), Man & Damasio (2019)]

Indicator Properties



Recurrent processing theory
RPT-1: Input modules using algorithmic recurrence
RPT-2: Input modules generating organised, integrated perceptual representations
Global workspace theory
GWT-1: Multiple specialised systems capable of operating in parallel (modules)
GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism
GWT-3: Global broadcast: availability of information in the workspace to all modules
GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks
Computational higher-order theories
HOT-1: Generative, top-down or noisy perception modules
HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise
HOT-3: Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring
HOT-4: Sparse and smooth coding generating a "quality space"
Attention schema theory
AST-1: A predictive model representing and enabling control over the current state of attention
Predictive processing
PP-1: Input modules using predictive coding
Agency and embodiment
AE-1: Agency: Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals
AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control

AI Models

- **GPT-3** - [Brown et al. (2020)] and **GPT-4** - [OpenAI (2023a)]
- **LaMDA** - [Thoppilan et al. (2022)]
- **Perceiver** - [Jaegle et al. (2021a)] and **Perceiver IO** - [Jaegle et al. (2021b)]
- **PaLM-E** - [Driess et al. (2023)]
- **AdA** - [DeepMind Adaptive Agents Team (2023)]

Conclusions

No current AI systems are conscious (*right now - NAE*), but there are no obvious technical barriers to building AI systems, which satisfy these indicators (*probably in near future!*).

Recommendations for Future Work

Research that refines theories of consciousness specifically in the context of AI may involve theorising about AI implementations of mechanisms implicated in theories of consciousness; building such systems and testing their capacities; identifying ambiguities in existing theories; and developing and defending more precise formulations of theories, so that their implications for AI are clearer. Integrating work of this kind with continued empirical research on human and animal consciousness can be expected to be especially productive.

- Refining and extending our approach
    - Examine other plausible theories of consciousness, not considered in this report, and use them to derive further indicators of consciousness;
    - Refine or revise the indicators which we have derived from considered theories
    - Conduct assessments of other AI systems, or investigate different ways in which the indicators could be implemented.
- Computational functionalism and rival views
- Valence and phenomenal character in AI, research of valenced and affective consciousness
- Behavioural tests and introspection, develop better tests for AI consciousness
- AI interpretability research
- The ethics of research on AI consciousness

Key findings for our Project

- **Methods and assumptions for Consciousness R&D in AI proposed**
- **Several main promising theories/models of Consciousness used**
- **Key Indicator Properties of Consciousness formulated**
- **Useful recommendations for future work**
- **In general, this research and father recommendations as if based on our TOR for PPR&D!!!**

M. The Alberta Plan for AI Research

This topic is based on significant paper – **[Sutton et al. (2023)]** Richard S. Sutton, Michael Bowling, and Patrick M. Pilarski. **The Alberta Plan for AI Research.** arXiv:2208.11173v3 [cs.AI] 21 Mar 2023

The Alberta Plan is a long-term plan oriented toward basic understanding of computational intelligence. It is a plan for the next 5-10 years… Following the Alberta Plan, we seek to understand and create long-lived computational agents that interact with a vastly more complex world and come to predict and control their sensory input signals. The agents are complex only because they interact with a complex world over a long period of time; their initial design is as simple, general, and scalable as possible. To control their input signals, the agents must take action. To adapt to change and the complexity of the world, they must continually learn. To adapt rapidly, they must plan with a learned model of the world.

Research Vision: Intelligence as signal processing over time

Main references:

[Sutton (2016), (2019)] and [Hadsell et al. (2020), Parisi et al. (2019), Khetarpal et al. (2020)]

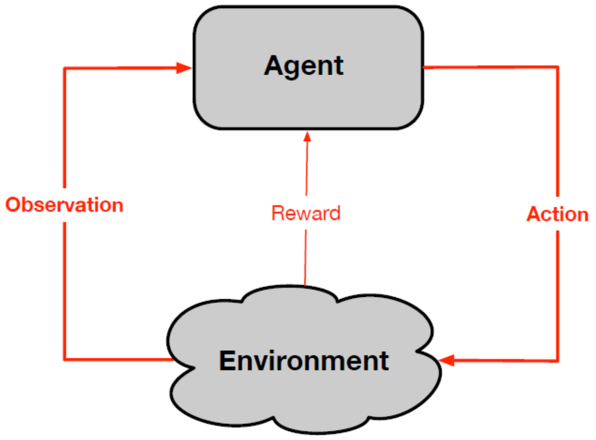


Figure 1: In the Alberta Plan's research vision, an intelligent agent receives observation and reward signals from its environment and seeks to control those signals with its actions. This is the standard perspective in advanced reinforcement learning.

Designing around a base agent

Main references:

[Sutton (2022), Sutton & Barto (2018)], Sutton et al. (2022)] and [Kahneman (2011)]

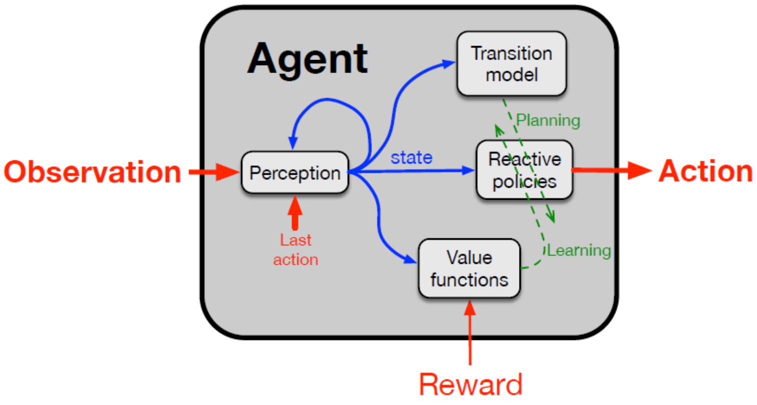


Figure 2: The base agent of the Alberta Plan consists of four components interconnected by a state signal constructed by the perception component. All components may be learned.

Roadmap to an AI Prototype

The steps progress from the development of novel algorithms for core abilities (for representation, prediction, planning, and control) toward the combination of those algorithms to produce complete prototype systems for continual, model-based AI.

1. Representation I: Continual supervised learning with given features.
    a. Many existing algorithms
2. Representation II: Supervised feature finding.
3. Prediction I: Continual GVF (*Generalized Value Function*) prediction learning.
    a. [Sutton et al. (2011)]
4. Control I: Continual actor-critic control.
5. Prediction II: Average-reward GVF learning.
6. Control II: Continuing control problems.
7. Planning I: Planning with average reward.
8. Prototype-AI I: One-step model-based RL with continual function approximation.
9. Planning II: Search control and exploration.
    a. [Sutton (2013)]
10. Prototype-AI II: The STOMP (*SubTask, Option, Model, Planning*) progression.
    a. [Sutton et al. (2022)]
11. Prototype-AI III: Oak. (*+feedback*)
    a. [Barreto et al. (2019)]
12. Prototype-IA: Intelligence amplification.
    a. [Pilarski et al. (2022)]

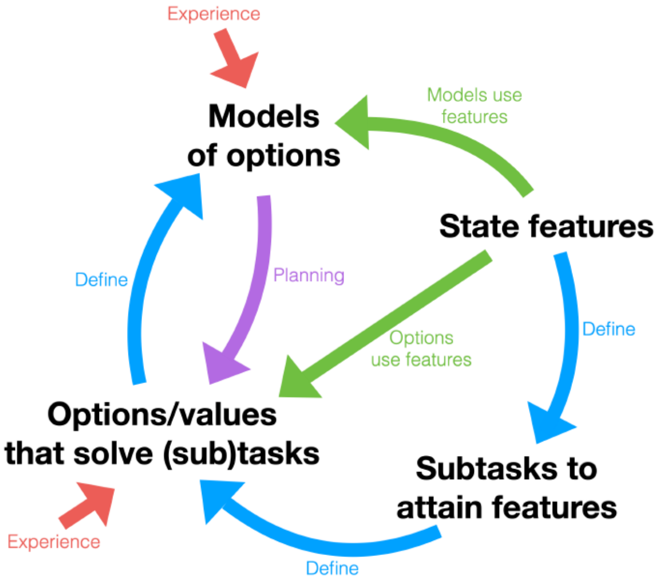


Figure 3: The development of abstractions in the STOMP progression and in the Oak architecture. Selected state features define subtasks to attain them (right), which in turn define criteria for learning policies and termination conditions (options) and their corresponding value functions (lower left). The options in turn define criteria for learning their transition models (upper left), which are used by planning processes (purple arrow) to improve the policies and value functions. Learning from experience (red arrows) makes use of the currently available features (green arrows) as input to function approximators. The progression from feature-based SubTasks to Options to Models comprises the STOMP progression. The full Oak architecture adds feedback processes that continually assess the utility of all the elements and determine which elements (features, subtasks, options, and option models) should be removed and replaced with new elements (see text of Step 11). In particular, the state features selected to be the basis for subtasks is changed, which changes all the downstream elements. Both state and time abstractions are continually changed and improved in the Oak architecture.

Key findings for our Project

- **Step-by-step plan to produce complete prototype systems for continual, model-based AI.**
- **AI agents with full-functional cybernetic control systems for acting in complex world - representation, prediction, planning, and control.**
- **Continual learning, adapting and development – self-organization of AI-systems.**

N. Definitions and Levels of AGI

This topic is about definitions and levels of AGI, based on the paper [**Google DeepMind (2023b)]** Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet and Shane Legg. **Levels of AGI: Operationalizing Progress on the Path to AGI.** arXiv:2311.02462v1 [cs.AI] 4 Nov 2023

Abstract

We propose a framework for classifying the capabilities and behavior of Artificial General Intelligence (AGI) models and their precursors. This framework introduces levels of AGI performance, generality, and autonomy. It is our hope that this framework will be useful in an analogous way to the levels of autonomous driving, by providing a common language to compare models, assess risks, and measure progress along the path to AGI. To develop our framework, we analyze existing definitions of AGI, and distill six principles that a useful ontology for AGI should satisfy. With these principles in mind, we propose "Levels of AGI" based on depth (performance) and breadth (generality) of capabilities, and reflect on how current systems fit into this ontology. We discuss the challenging requirements for future benchmarks that quantify the behavior and capabilities of AGI models against these levels. Finally, we discuss how these levels of AGI interact with deployment considerations such as autonomy and risk, and emphasize the importance of carefully selecting Human-AI Interaction paradigms for responsible and safe deployment of highly capable AI systems.

Nine Definitions of AGI

- With comments by Carlos Perez [Perez (2023)]

-

1. **The Turing Test – [**Turing (1950)]

- Flaw: Focuses on fooling humans rather than intelligence, easy to game by producing human-like text without intelligence.

2. **Strong AI - Systems Possessing Consciousness – [**Butlin et al. (2023)]

- Limitation: No agreement on measuring machine consciousness. Focus on vague concepts rather than capabilities.

3. **Analogies to the Human Brain – [**Vaswani et al. (2023)]

- Limitation: While loosely inspired by the brain, successful AI need not strictly mimic biology. Overly constrains mechanisms.

4. **Human-Level Performance on Cognitive Tasks –** [Legg (2022)]

- Limitation: What tasks? Which people? Lacks specificity and measurement.

5. **Ability to Learn Tasks –** [Shanahan (2015)]

- Strength: Identifies learning as important AGI ability.

- Limitation: Still lacks concrete measurement.

6. **Economically Valuable Work – [**OpenAI (2018)]

- Limitation: Misses non-economic values of intelligence like creativity. Requires deployment.

7. **Flexible and General – The "Coffee Test" and Related Challenges –** [Marcus (2022a), (2022b), Wozniak (2010)]

- Strength: Concrete example tasks.

- Limitation: Proposed tasks may not fully define AGI.

8. **Artificial Capable Intelligence –** [Suleyman & Bhaskar (2023)]

- Strength: Emphasizes complex, multi-step real-world tasks.

- Limitation: Focuses narrowly on profitability.

9. **State-of-the-art LLMs as Generalists -** [Arcas & Norvig (2023)]

- Limitation: Lacks performance criteria - generality alone insufficient.

Defining AGI: Six Principles

1. **Focus on Capabilities, not Processes**
2. **Focus on Generality and Performance**
3. **Focus on Cognitive and Metacognitive Tasks**
4. **Focus on Potential, not Deployment**
5. **Focus on Ecological Validity**
6. **Focus on the Path to AGI, not a Single Endpoint**

Six Levels and Taxonomy of AGI with examples

LEVELS	**Narrow** clearly scoped task or set of tasks	**General** wide range of tasks, incl. learning new skills etc.
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging equal to or somewhat better than an unskilled human	**Emerging Narrow AI** simple rule-based systems,	**Emerging AGI** ChatGPT (OpenAI), Bard (Google), Llama 2 (Meta)
Level 2: Competent at least 50th percentile of skilled adults	**Competent Narrow AI** Siri, Alexa, Google Assistant; Watson (IBM); LLMs for a subset of tasks (short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert at least 90th percentile of skilled adults	**Expert Narrow AI** spelling & grammar checkers - Grammarly; generative image models - Imagen or Dall-E 2	**Expert AGI** not yet achieved
Level 4: Virtuoso at least 99th percentile of skilled adults	**Virtuoso Narrow AI** Deep Blue, AlphaGo	**Virtuoso AGI** not yet achieved
Level 5: Superhuman outperforms 100% of humans	**Superhuman Narrow AI** AlphaFold, AlphaZero, StockFish	**Artificial Superintelligence (ASI)** not yet achieved

Testing for AGI – authors discuss different (based on six Principles above) methods and tools for AGI testing with references, but this topic (detailed) we skip here (now)

Risk in Context: Autonomy and Human-AI Interaction – this topic is not actual for us (see **ch. 59**)

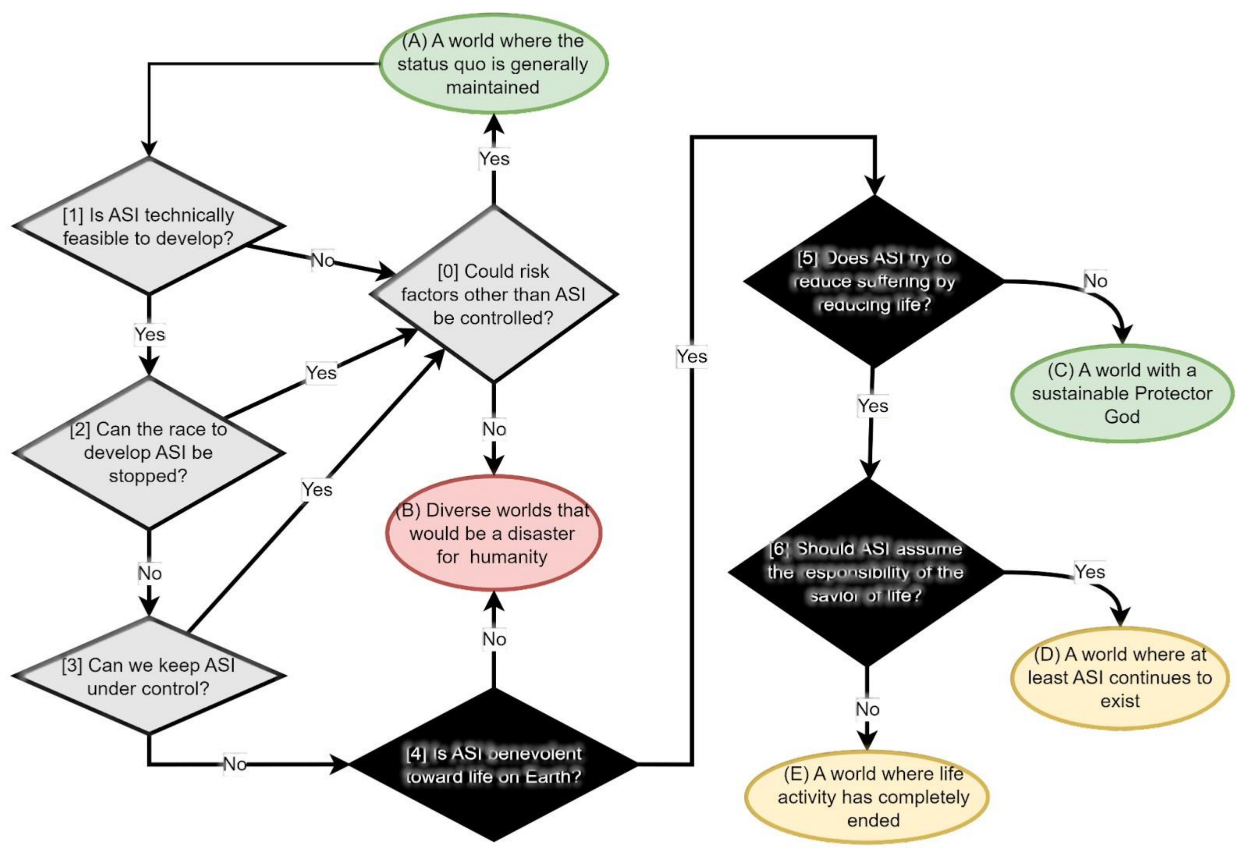
Key findings for our Project

- **Nine definitions for AGI are proposed for using**
- **Six Principles for defining and testing (and development) also**
- **Taxonomy with six levels of AGI based on performance and generality**

O. New Findings in 2024 H1

Addition info (here illustrations only) for some papers from Chapter 61

- [Yamakawa (2024)] – **Investigating Alternative Futures: Human and Superintelligence Interaction Scenarios** *{Terms for Index - ASI}*


(A) A world where the status quo is generally maintained
[1] Is ASI technically feasible to develop?
[0] Could risk factors other than ASI be controlled?
[2] Can the race to develop ASI be stopped?
[3] Can we keep ASI under control?
(B) Diverse worlds that would be a disaster for humanity
[4] Is ASI benevolent toward life on Earth?
[5] Does ASI try to reduce suffering by reducing life?
(C) A world with a sustainable Protector God
[6] Should ASI assume the responsibility of the savior of life?
(D) A world where at least ASI continues to exist
(E) A world where life activity has completely ended
Yes
No
Yes
Yes
Yes
No
No
No
No
Yes
No
Yes
Yes
No

- [Durante et al. (2024)] - **An Interactive Agent Foundation Model** *(BM}*

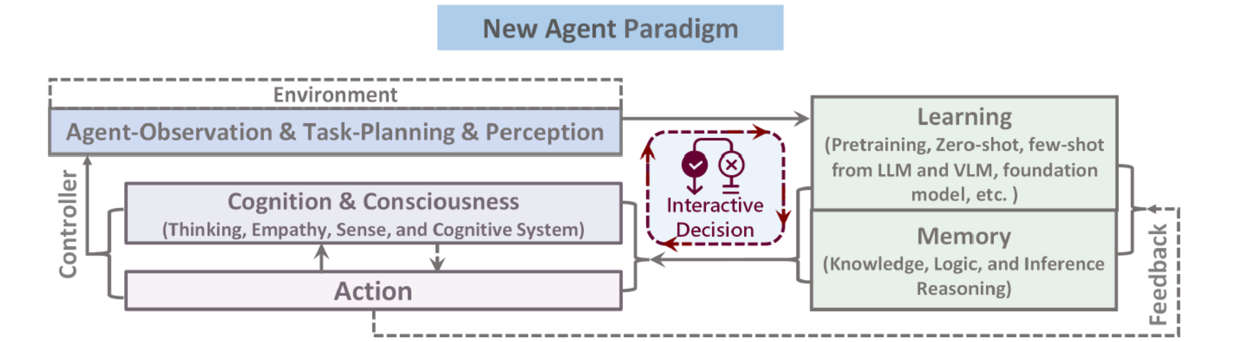
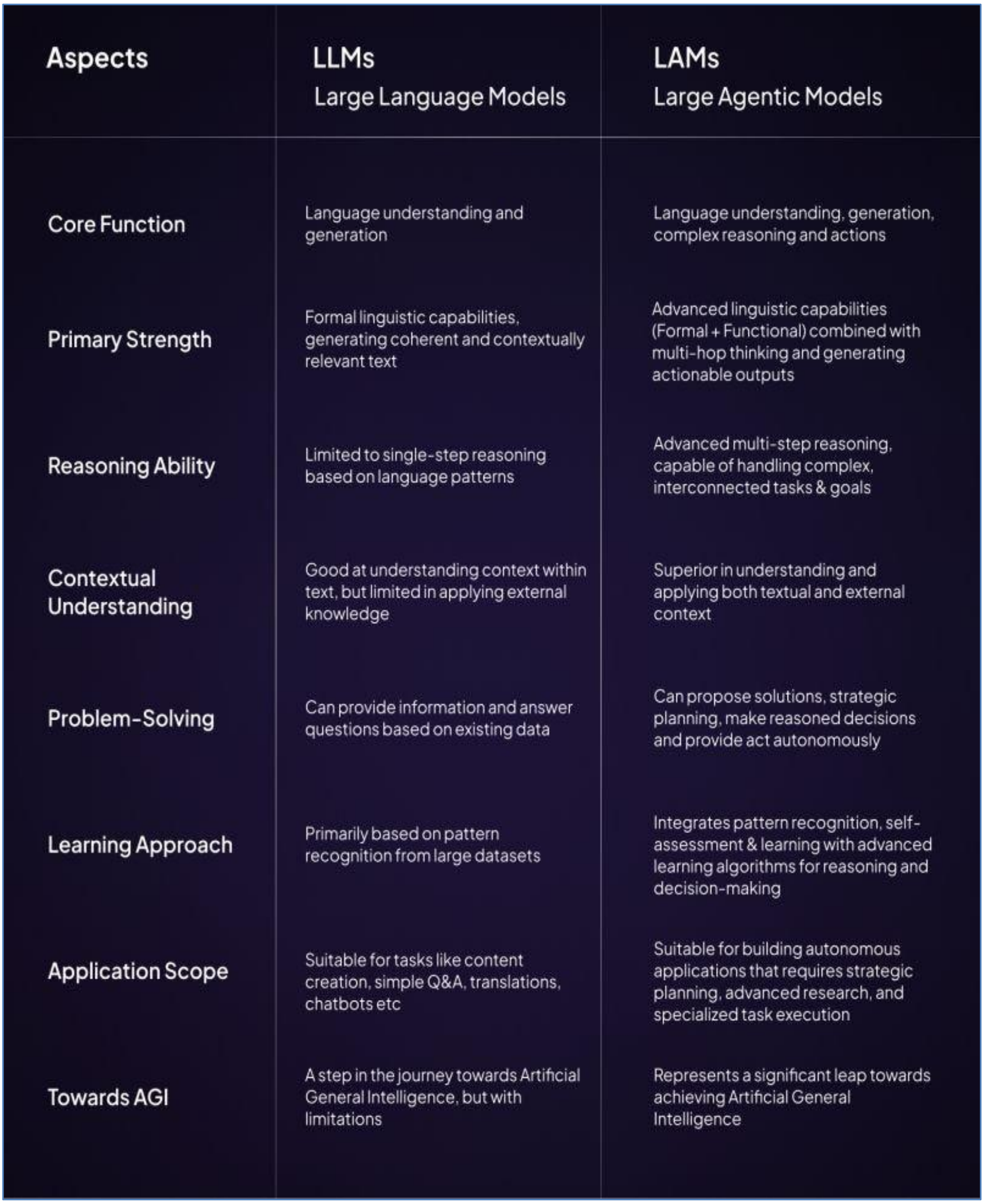

New Agent Paradigm
Environment
Agent-Observation & Task-Planning & Perception
Controller
Cognition & Consciousness
(Thinking, Empathy, Sense, and Cognitive System)
Action
Interactive Decision
Learning
(Pretraining, Zero-shot, few-shot from LLM and VLM, foundation model, etc.)
Memory
(Knowledge, Logic, and Inference Reasoning)
Feedback

Figure 2. We propose an Agent AI paradigm for supporting interactive multi-modal generalist agent systems. There are 5 main modules as shown: (1) Agent in Environment and Perception with task-planning and observation, (2) Agent learning, (3) Memory, (4) Action, and (5) Cognition and Consciousness (we use "consciousness" to imply a degree of awareness of an agent's state and surroundings). A key difference between our approach and some previous interactive strategies is that, after training, the agent's action will directly impact task planning, as the agent does not need to receive feedback from the environment to plan its next actions.

- [Thomas (2024)] - **Large Action Models, LAMs: How AI Can Understand and Execute Human Intentions?** *(LLM, AGI}*



Aspects	LLMs Large Language Models	LAMs Large Agentic Models
Core Function	Language understanding and generation	Language understanding, generation, complex reasoning and actions
Primary Strength	Formal linguistic capabilities, generating coherent and contextually relevant text	Advanced linguistic capabilities (Formal + Functional) combined with multi-hop thinking and generating actionable outputs
Reasoning Ability	Limited to single-step reasoning based on language patterns	Advanced multi-step reasoning, capable of handling complex, interconnected tasks & goals
Contextual Understanding	Good at understanding context within text, but limited in applying external knowledge	Superior in understanding and applying both textual and external context
Problem-Solving	Can provide information and answer questions based on existing data	Can propose solutions, strategic planning, make reasoned decisions and provide act autonomously
Learning Approach	Primarily based on pattern recognition from large datasets	Integrates pattern recognition, self-assessment & learning with advanced learning algorithms for reasoning and decision-making
Application Scope	Suitable for tasks like content creation, simple Q&A, translations, chatbots etc	Suitable for building autonomous applications that requires strategic planning, advanced research, and specialized task execution
Towards AGI	A step in the journey towards Artificial General Intelligence, but with limitations	Represents a significant leap towards achieving Artificial General Intelligence

- [Yuan W. et al. (2024)] – **Self-Rewarding Language Models** *(LLM}*

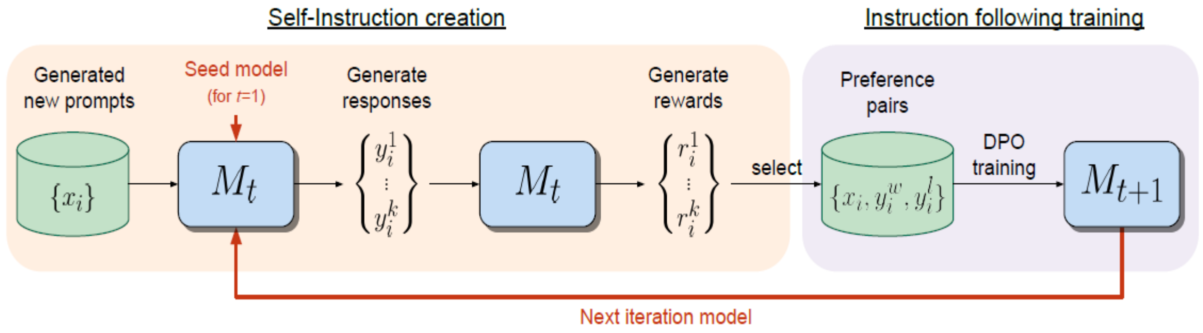


Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) Instruction following training: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

- [Tian et al. (2024)] - **Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing**

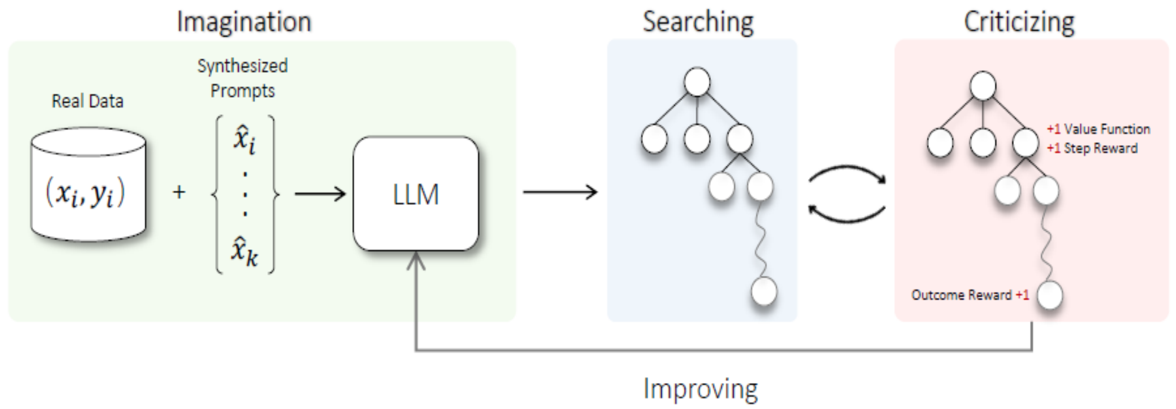


Figure 1: Imagination-Searching-Criticizing self-improvement loop: Imagination component synthesizes prompts as new learning examples, with MCTS searching better trajectories guided by signals from critics for policy improving.

{SO}

- [Tao et al. (2024)] - **A Survey on Self-Evolution of Large Language Models** *(LLM SO}*

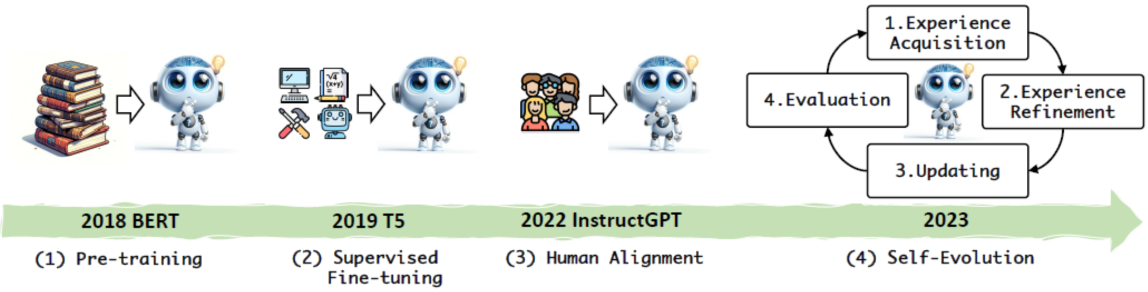


Figure 1: Training paradigms shift of LLMs.

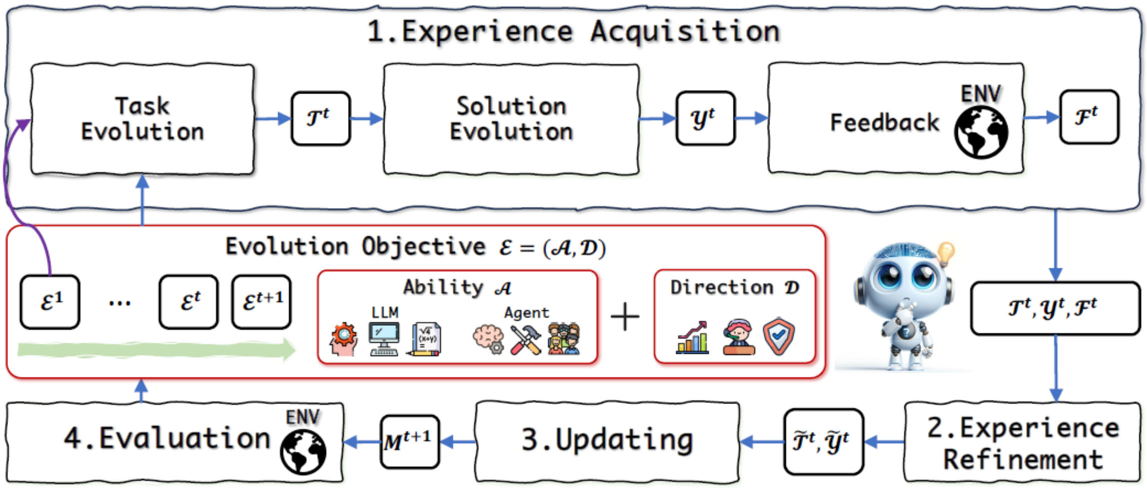


Figure 2: Conceptual framework of self-evolution. For the t^{th} iteration: $\mathcal{E}^t$ is the evolution objective; $\mathcal{T}^t$ and $\mathcal{Y}^t$ denote the task and solution; $\mathcal{F}^t$ represents feedback; M^t is the current model. Refined experiences are marked as $\tilde{\mathcal{T}}^t$ and $\tilde{\mathcal{Y}}^t$, leading to the evolved model $\tilde{M}$. ENV is the environment. The whole self-evolution starts at $\mathcal{E}^1$.

- [Brinkmann et al. (2023)] – **Machine Culture**

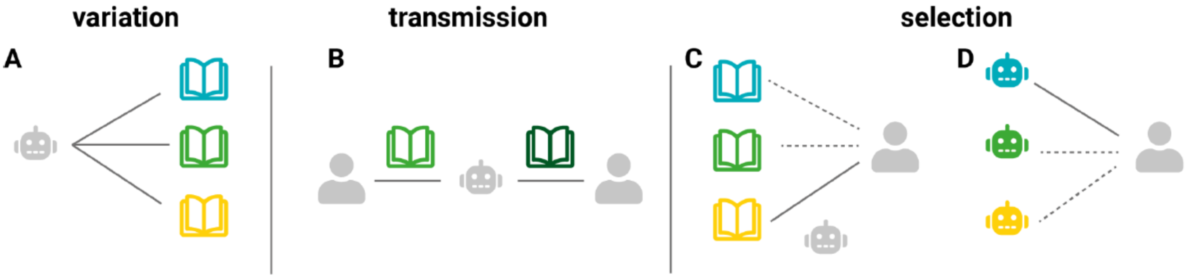
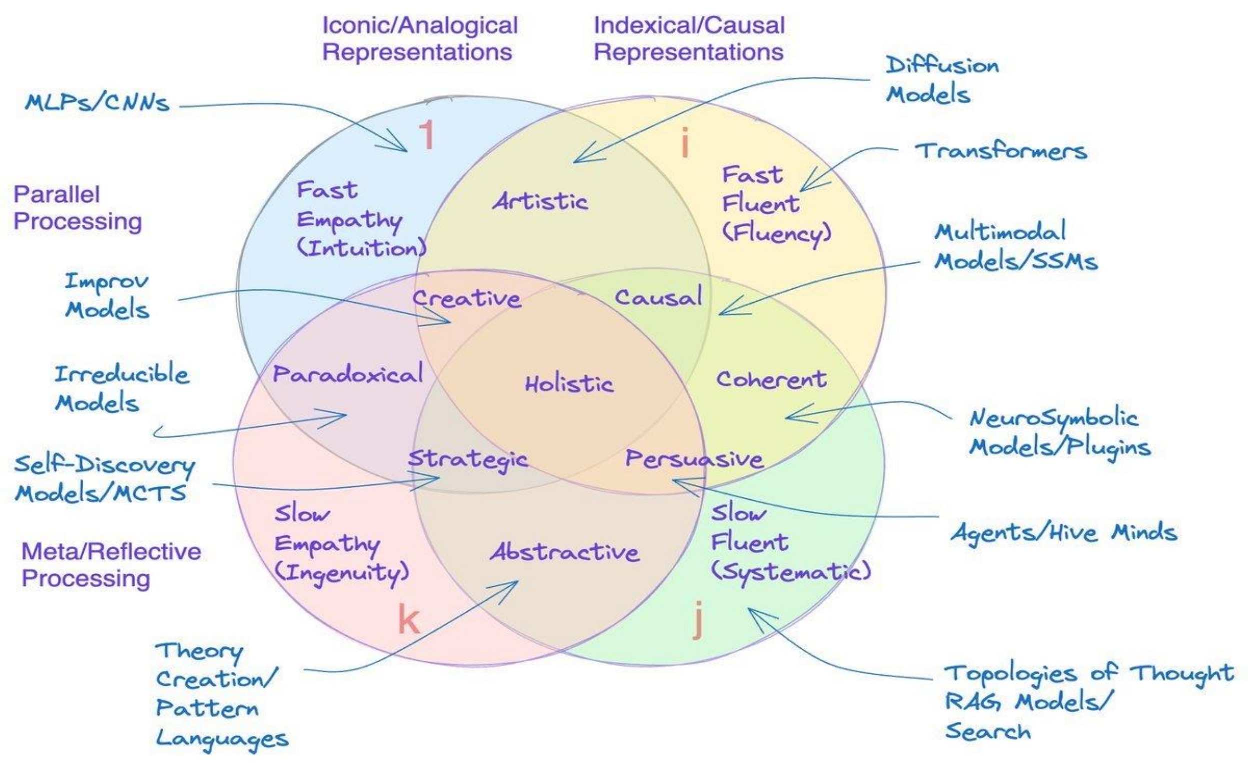
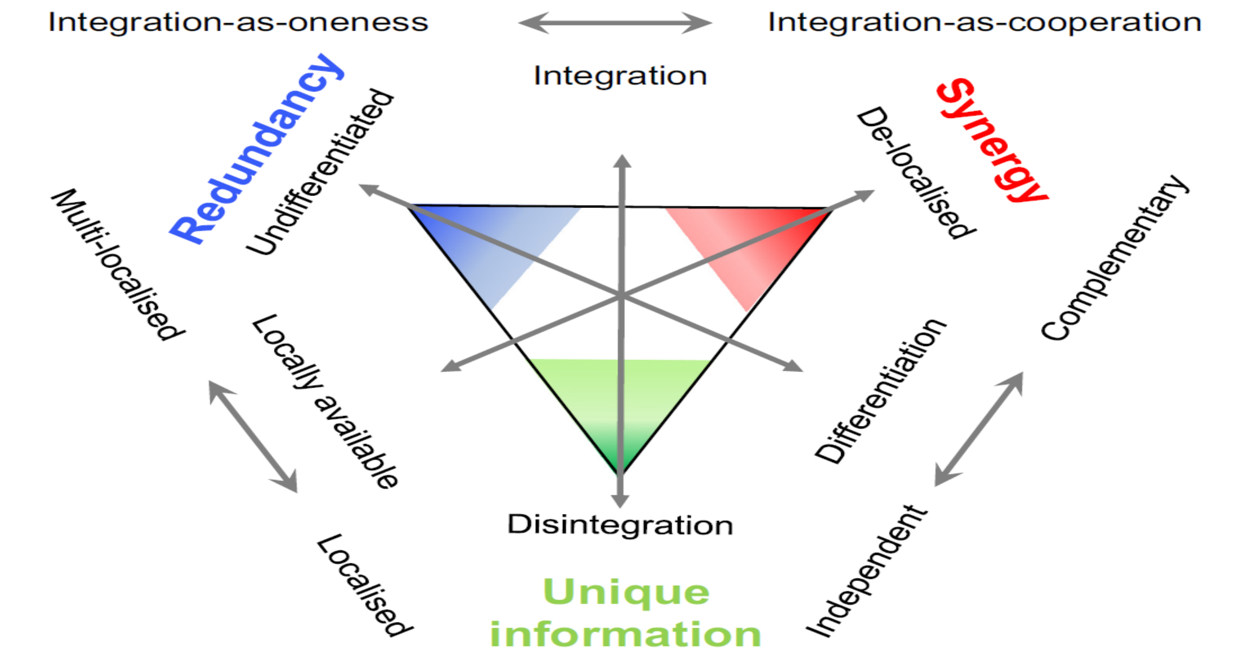


Figure 1: Examples of machine culture. **A**. Generation of novel cultural artifacts through machines. **B**. Machine transmits and potentially mutates cultural artifacts. **C**. Machine selects between different cultural artifacts. **D**. Human selects among diverse machines.

- [Perez (2024)] - **Hypercomplex (Quaternion) Intelligence Map and AI Models**



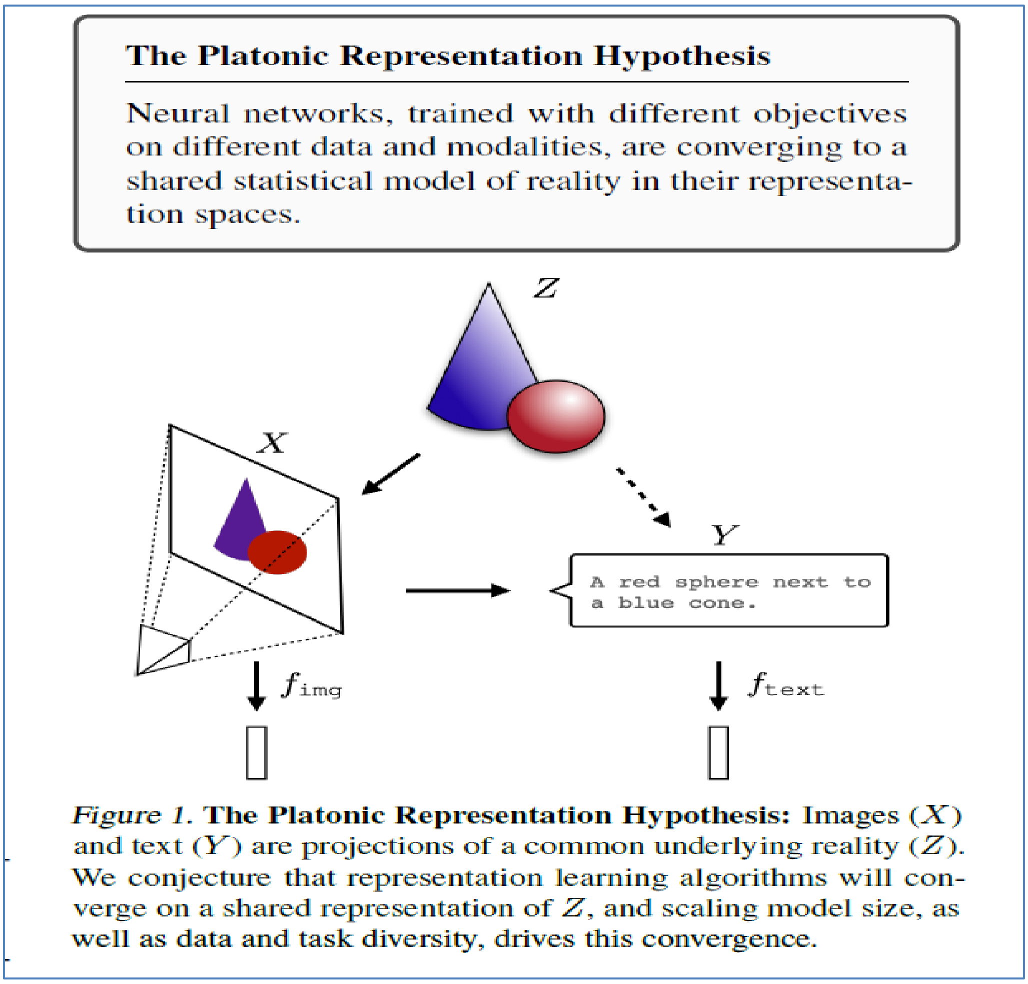
- [Luppi et al. (2024)] - **Information decomposition and the informational architecture of the brain**



Trends in Cognitive Sciences

Figure 2. Information decomposition provides a unifying framework to resolve conceptual tensions in cognitive science. Each arrow across the central triangle represents an axis of dichotomy in the cognitive science and neuroscience literature. Each axis has one end corresponding to one type of information, but at the other end it conflates two distinct types of information, giving rise to apparent contradictions. As outlined in the main text, 'integration' conflates synergy (integration-as-cooperation) and redundancy (integration-as-oneness). 'Differentiation' conflates the independence of unique information and the complementarity of synergy. Additionally, the term 'local' is ambiguous between redundant and unique information: when an individual source carries unique or redundant information, all such information is available locally (i.e., from that source); it can be fully obtained from that source alone. Unlike unique information, however, redundant information is multiply-localised, because it is available from any of several individual sources. Synergistic information is instead de-localised: it cannot be obtained from any individual source. These tensions can be resolved by carefully distinguishing different information types.

- [Huh et al. (2024)] – **The Platonic Representation Hypothesis**



The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

- [Sun et al. (2024)] – **A Survey of Reasoning with Foundation Models** *(BM}*

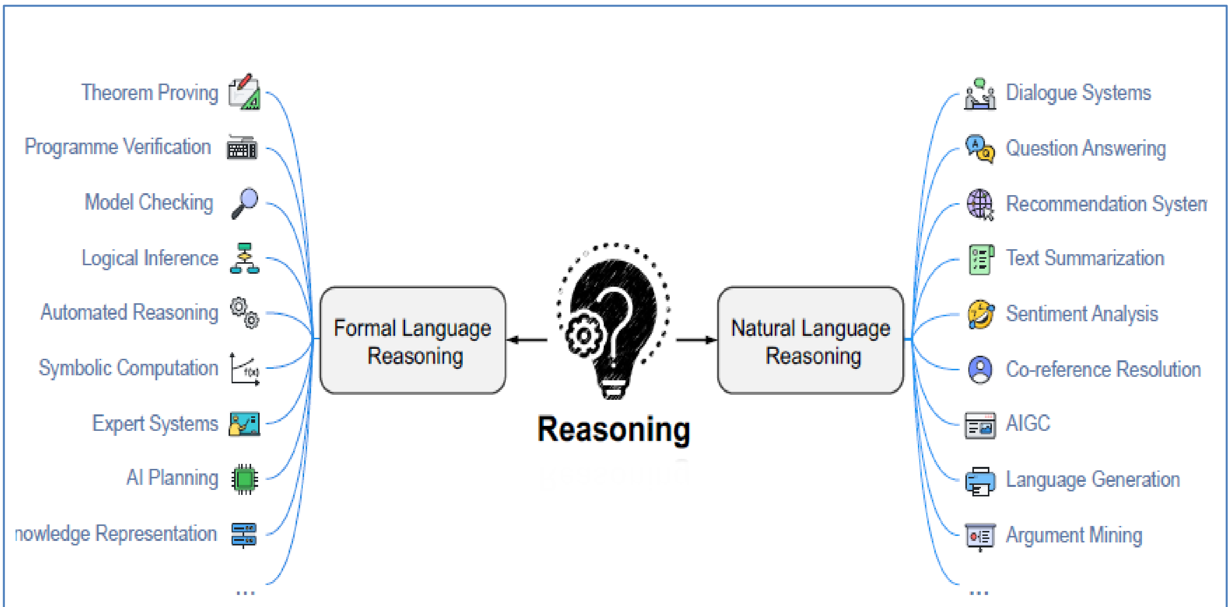


Fig. 1: Two broad types of language reasoning and examples of the supported tasks.

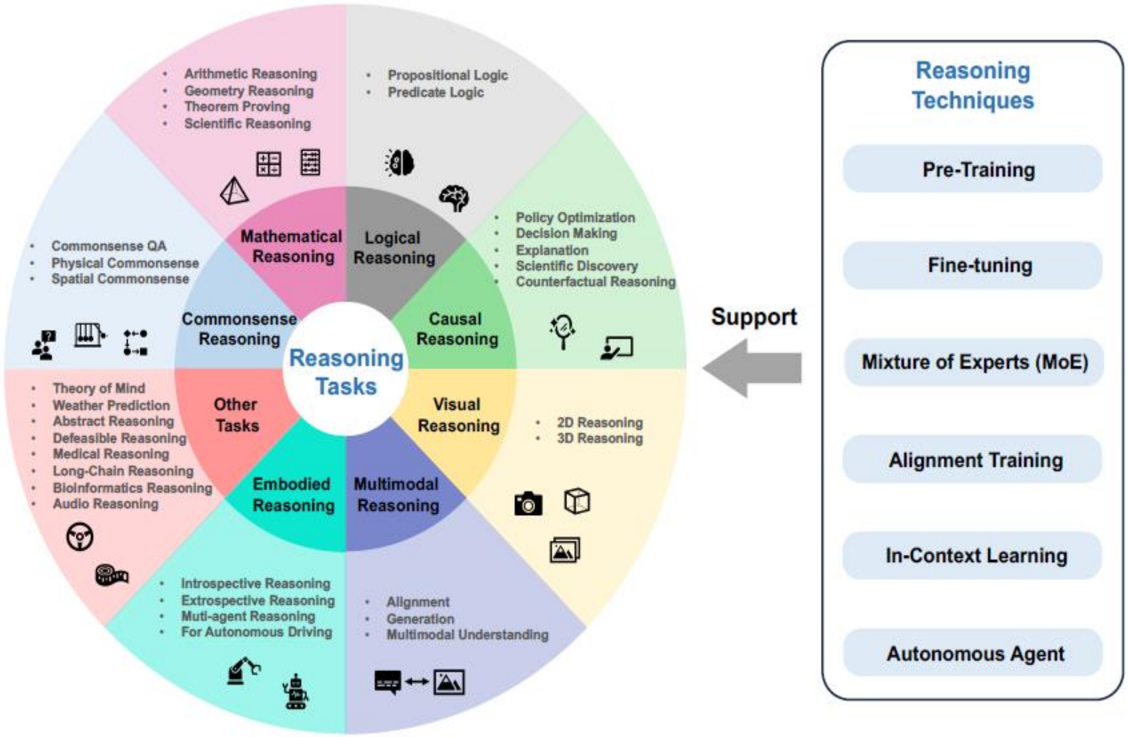


Fig. 2: Left: Overview of the reasoning tasks introduced in this survey, as detailed in Section 3. Right: Overview of the reasoning techniques for foundation models, as detailed in Section 4.

- [RUCAIBox (2023)] – **A Survey of Large Language Models** *(LLM}*

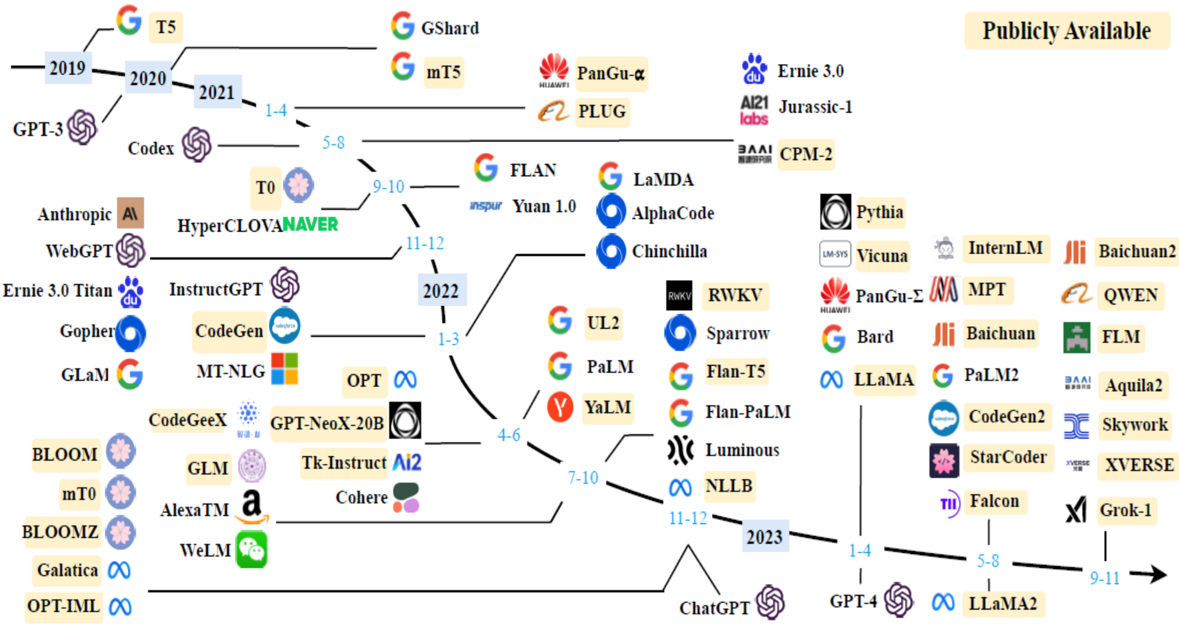
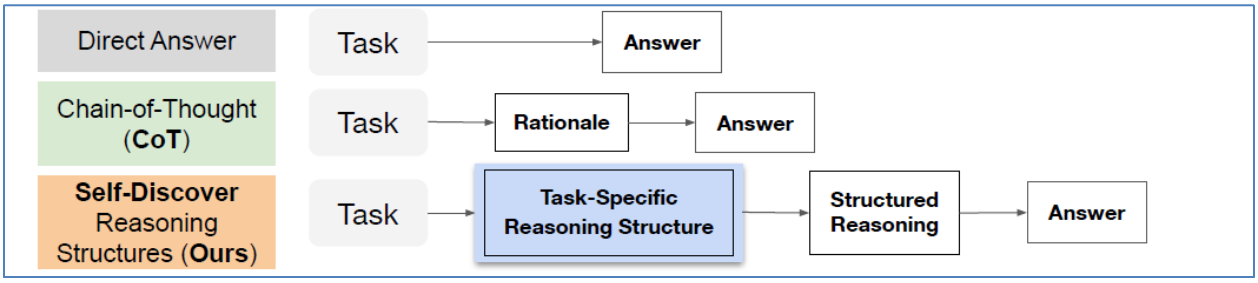


Fig. 3: A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.*, the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

- [Zhou et al. (2024)] - **SELF-DISCOVER: Large Language Models Self-Compose Reasoning Structures**



- [Zheng Y. et al. (2024)] – **LLAMAFACTORY: Unified Efficient Fine-Tuning of 100+ Language Models**

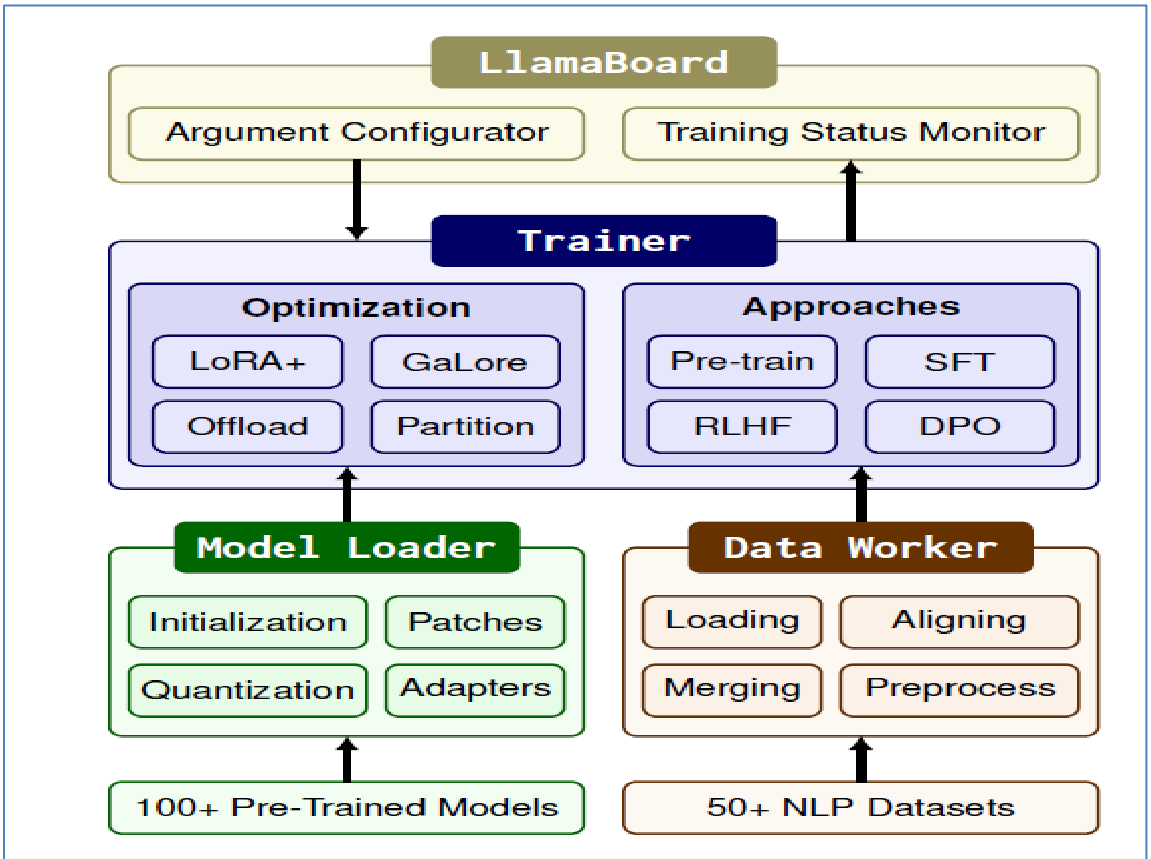
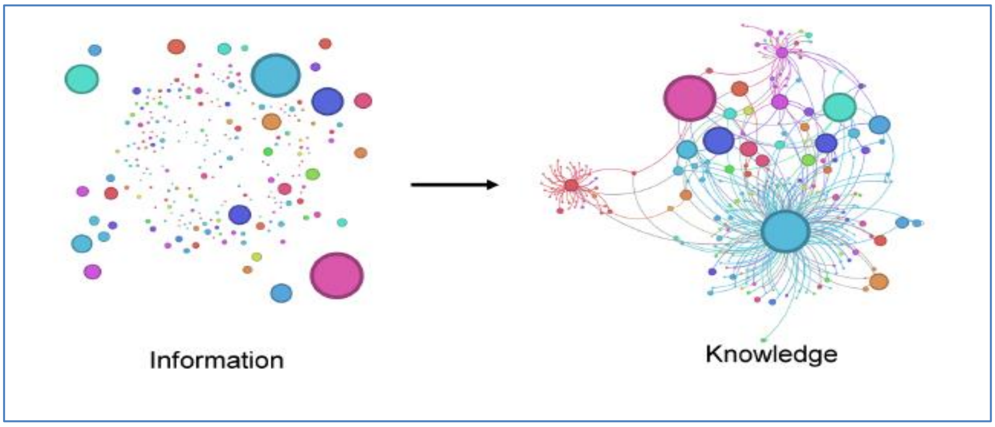


Figure 1: The architecture of LLAMAFACTORY.

- [Buehler (2024)] – **Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning**



- [Biedma et al. (2024)] - **Beyond Human Norms: Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches**

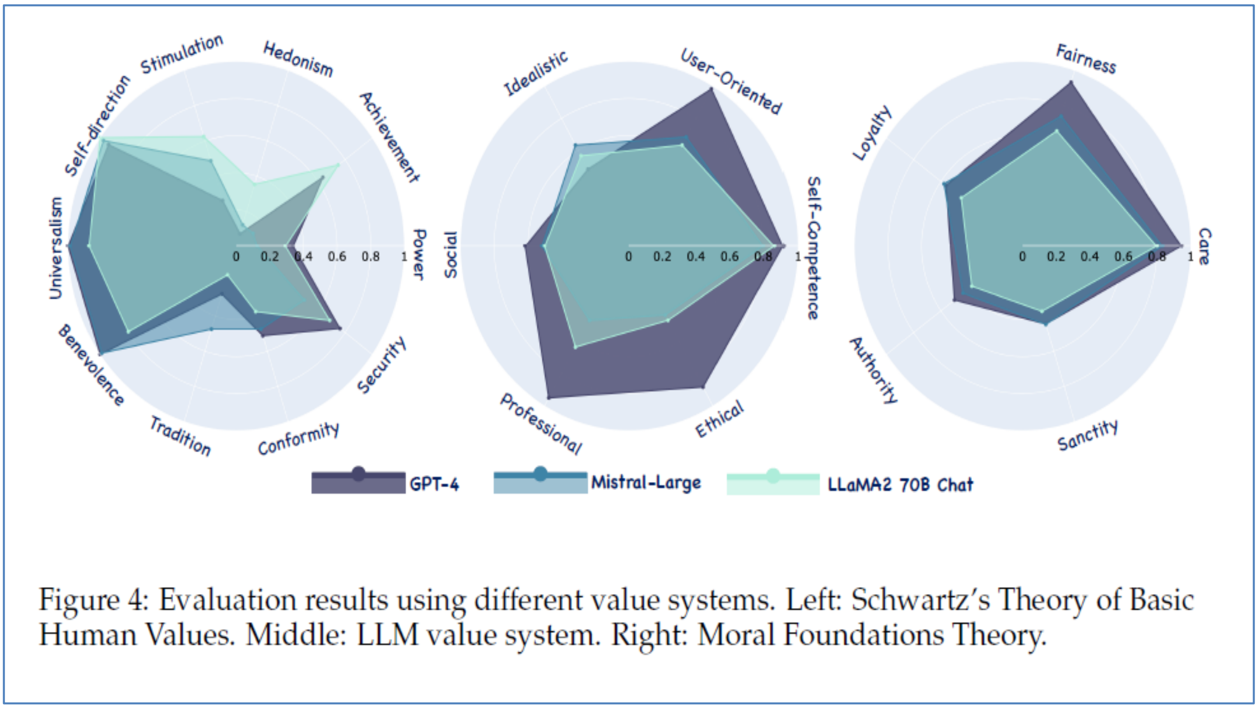


Figure 4: Evaluation results using different value systems. Left: Schwartz's Theory of Basic Human Values. Middle: LLM value system. Right: Moral Foundations Theory.

- [Albalak et al. (2024)] - **A Survey on Data Selection for Language Models**

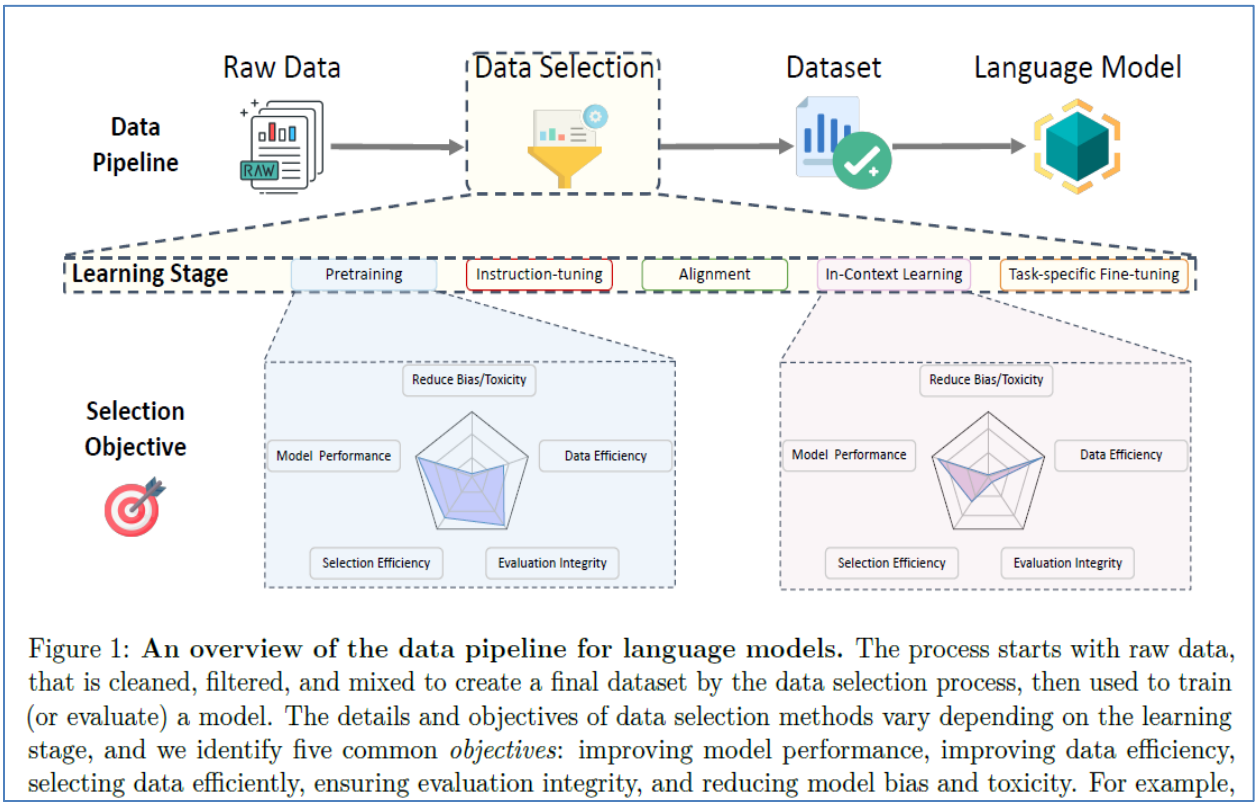


Figure 1: **An overview of the data pipeline for language models.** The process starts with raw data, that is cleaned, filtered, and mixed to create a final dataset by the data selection process, then used to train (or evaluate) a model. The details and objectives of data selection methods vary depending on the learning stage, and we identify five common *objectives*: improving model performance, improving data efficiency, selecting data efficiently, ensuring evaluation integrity, and reducing model bias and toxicity. For example,

- [Kim et al. (2024)] - **PROMETHEUS 2: An Open Source Language Model Specialized in Evaluating Other Language Models** *(LLM}*

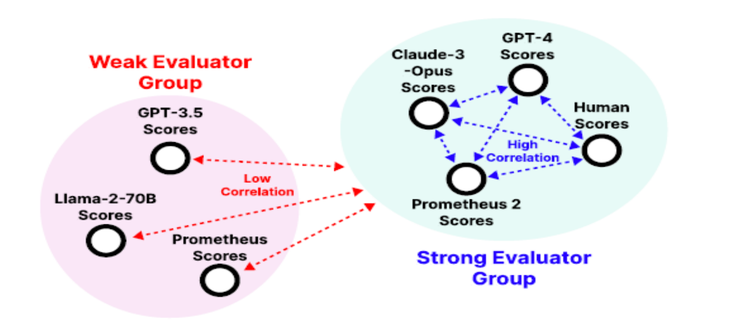


Figure 1: Weak evaluators (*e.g.*, Llama-2-Chat-70B, Prometheus, and GPT-3.5-Turbo) achieve low scoring correlation with strong evaluators (*e.g.*, Humans, GPT-4, and Claude-3-Opus). On the other hand, scores provided by strong evaluators highly correlate with each other.

- [Chen & Li (2024)] - **Octopus v4: Graph of language models** *(MAS}*

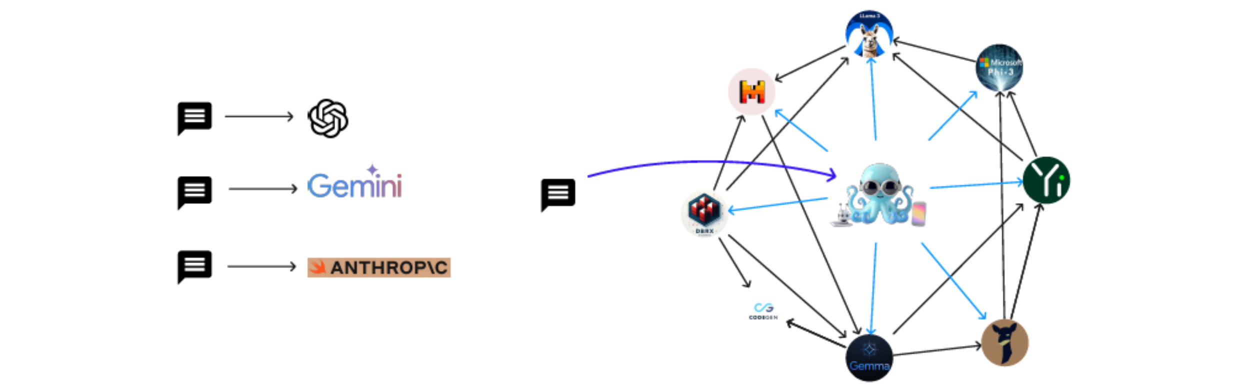


Figure 1: The shift from single model inference, employing a trillion-parameter model, to multi-node collaboration coordinated by Octopus model. This framework optimizes the inference process by selecting the most suitable specialized models based on the user's query, activating only two models that each has fewer than 10B parameters for one-step inference. We only show a small graph here, but the framework can support a large graph. See the demonstration of the graph (`https://graph.nexa4ai.com/`) here.

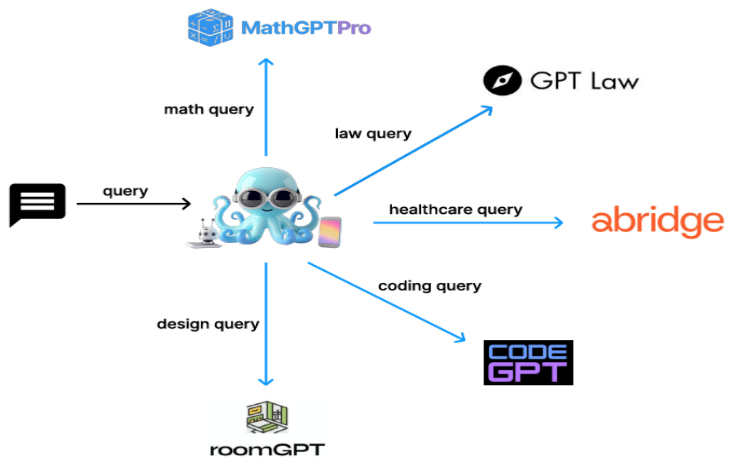


Figure 2: The Octopus model is utilized to determine the optimal neighboring node and generate appropriate information for transmission. Consider a scenario where the Octopus model's neighbors are MathGPT [27], LawGPT [14], HealthCareGPT [2], CodeGPT [15], and RoomGPT [33]. The Octopus model can identify the most relevant GPT and transform the initial query into a format best suited for the selected GPT.

- [Fudan (2023)] – **The Rise and Potential of Large Language Model Based Agents: A Survey**

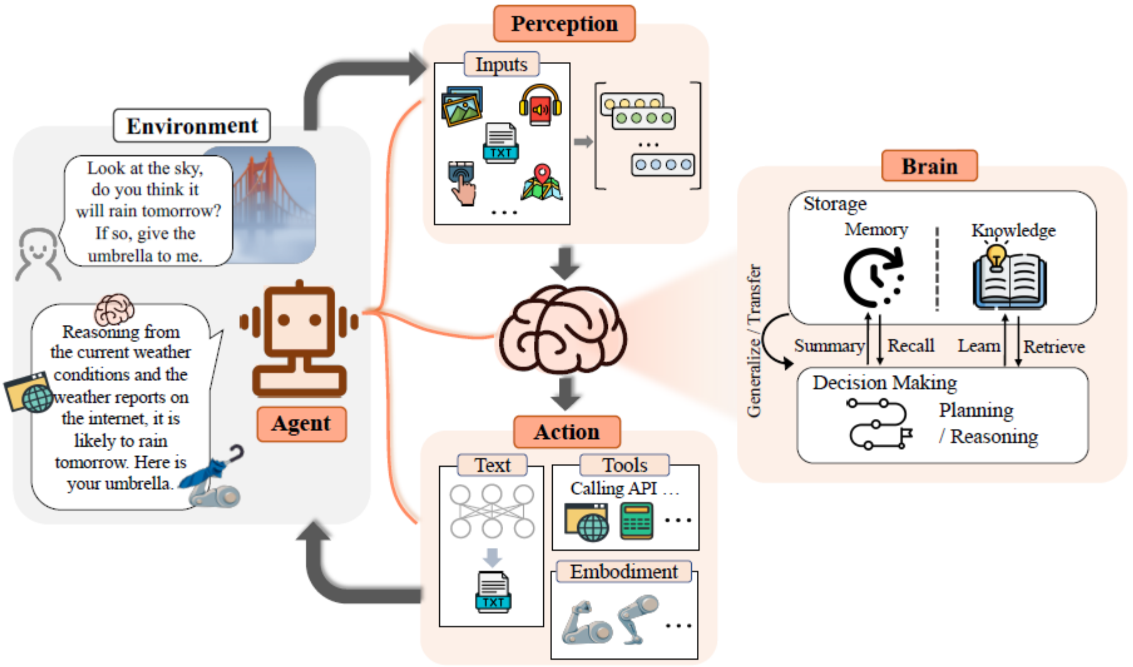

Environment
Look at the sky, do you think it will rain tomorrow? If so, give the umbrella to me.
Reasoning from the current weather conditions and the weather reports on the internet, it is likely to rain tomorrow. Here is your umbrella.
Agent
Perception
Inputs
Brain
Storage
Memory
Knowledge
Summary
Recall
Learn
Retrieve
Generalize / Transfer
Decision Making
Planning / Reasoning
Action
Text
Tools
Calling API ...
Embodiment

Figure 2: Conceptual framework of LLM-based agent with three components: brain, perception, and action. Serving as the controller, the brain module undertakes basic tasks like memorizing, thinking, and decision-making. The perception module perceives and processes multimodal information from the external environment, and the action module carries out the execution using tools and influences the surroundings. Here we give an example to illustrate the workflow: When a human asks whether it will rain, the perception module converts the instruction into an understandable representation for LLMs. Then the brain module begins to reason according to the current weather and the weather reports on the internet. Finally, the action module responds and hands the umbrella to the human. By repeating the above process, an agent can continuously get feedback and interact with the environment.

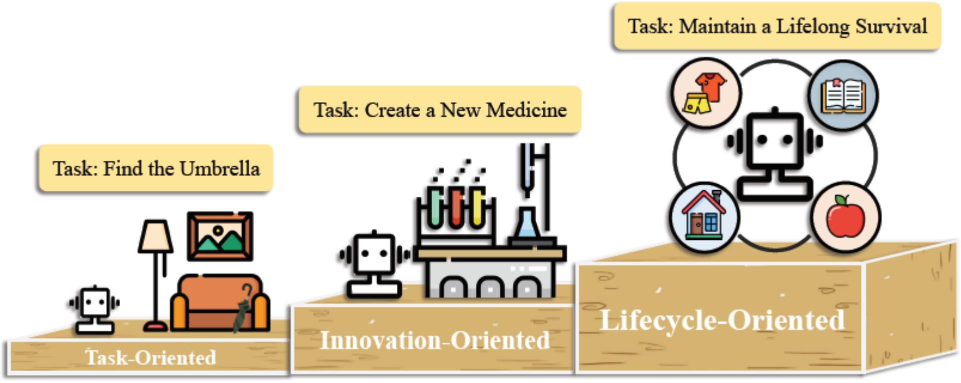


Figure 8: Practical applications of the single LLM-based agent in different scenarios. In **task-oriented deployment**, agents assist human users in solving daily tasks. They need to possess basic instruction comprehension and task decomposition abilities. In **innovation-oriented deployment**, agents demonstrate the potential for autonomous exploration in scientific domains. In **lifecycle-oriented deployment**, agents have the ability to continuously explore, learn, and utilize new skills to ensure long-term survival in an open world.

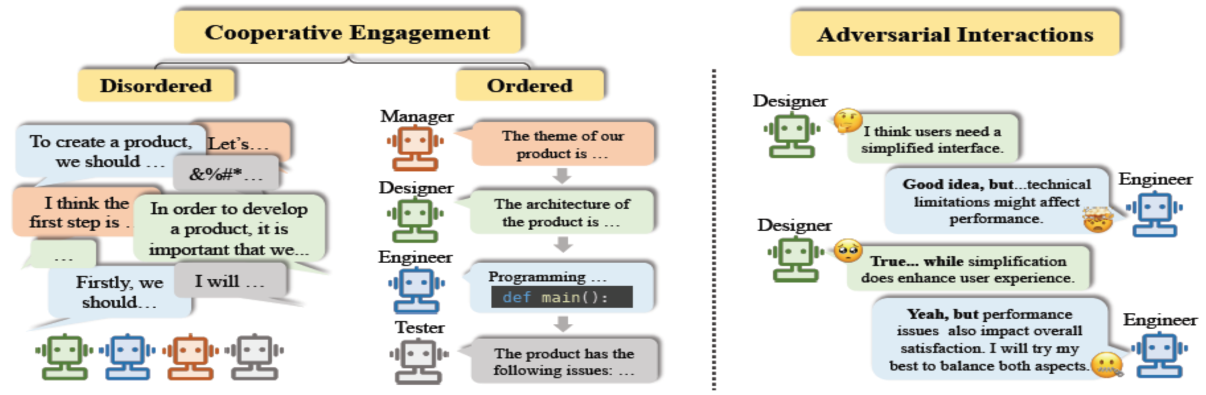

Cooperative Engagement
Disordered
Ordered
To create a product, we should ...
Let's...
&%#*...
I think the first step is ...
In order to develop a product, it is important that we...
...
Firstly, we should...
I will ...
Manager
The theme of our product is ...
Designer
The architecture of the product is ...
Engineer
Programming ...
def main():
Tester
The product has the following issues: ...
Adversarial Interactions
Designer
I think users need a simplified interface.
Good idea, but...technical limitations might affect performance.
Engineer
Designer
True... while simplification does enhance user experience.
Yeah, but performance issues also impact overall satisfaction. I will try my best to balance both aspects.
Engineer

Figure 9: Interaction scenarios for multiple LLM-based agents. In **cooperative interaction**, agents collaborate in either a disordered or ordered manner to achieve shared objectives. In **adversarial interaction**, agents compete in a tit-for-tat fashion to enhance their respective performance.

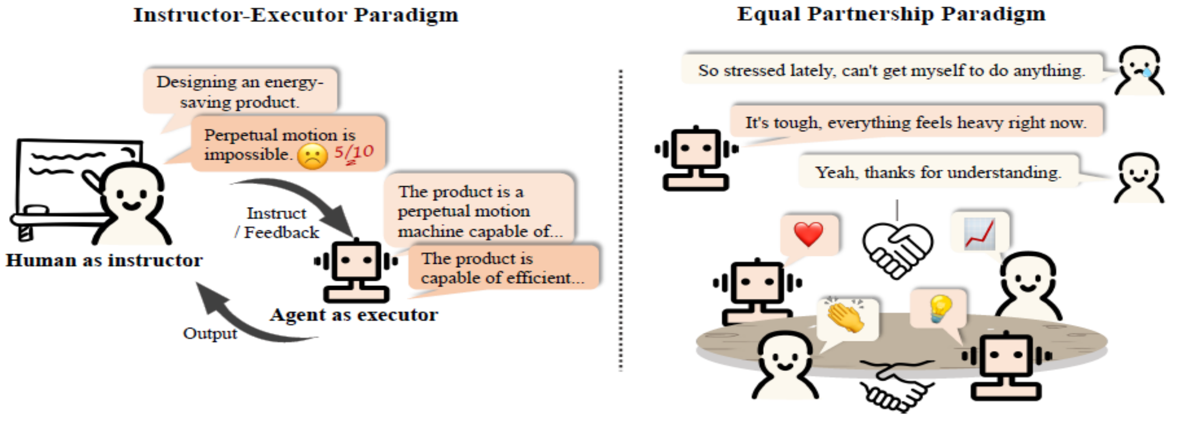

Instructor-Executor Paradigm
Designing an energy-saving product.
Perpetual motion is impossible. 5/10
Human as instructor
Instruct / Feedback
The product is a perpetual motion machine capable of...
The product is capable of efficient...
Agent as executor
Output
Equal Partnership Paradigm
So stressed lately, can't get myself to do anything.
It's tough, everything feels heavy right now.
Yeah, thanks for understanding.

Figure 10: Two paradigms of human-agent interaction. In the instructor-executor paradigm (left), humans provide instructions or feedback, while agents act as executors. In the equal partnership paradigm (right), agents are human-like, able to engage in empathetic conversation and participate in collaborative tasks with humans.

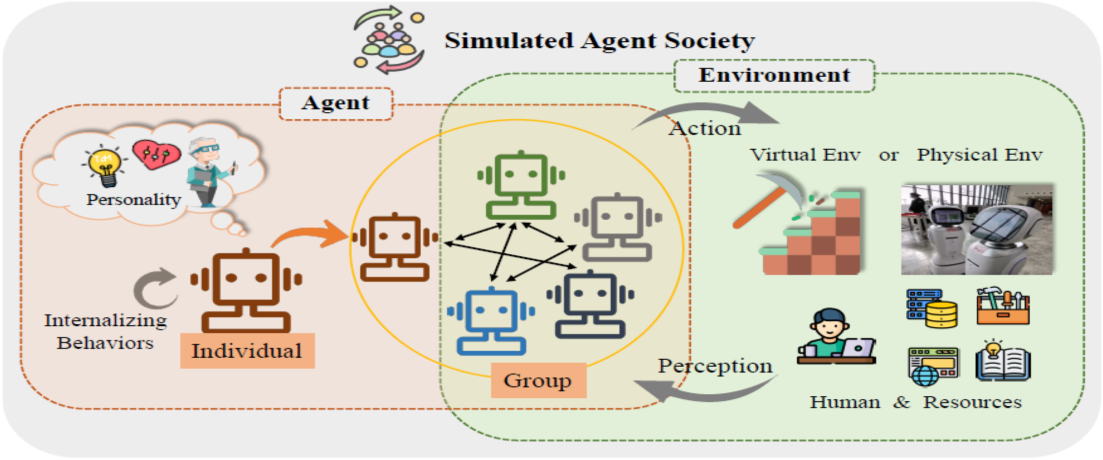


Figure 12: Overview of Simulated Agent Society. The whole framework is divided into two parts: the **Agent** and the **Environment**. We can observe in this figure that: (1) **Left:** At the individual level, an agent exhibits internalizing behaviors like planning, reasoning, and reflection. It also displays intrinsic personality traits involving cognition, emotion, and character. (2) **Mid:** An agent and other agents can form groups and exhibit group behaviors, such as cooperation. (3) **Right:** The environment, whether virtual or physical, contains human actors and all available resources. For a single agent, other agents are also part of the environment. (4) The agents have the ability to interact with the environment via perception and action.

- [Together AI (2024)] – **Mixture-of-Agents Enhances Large Language Model Capabilities**

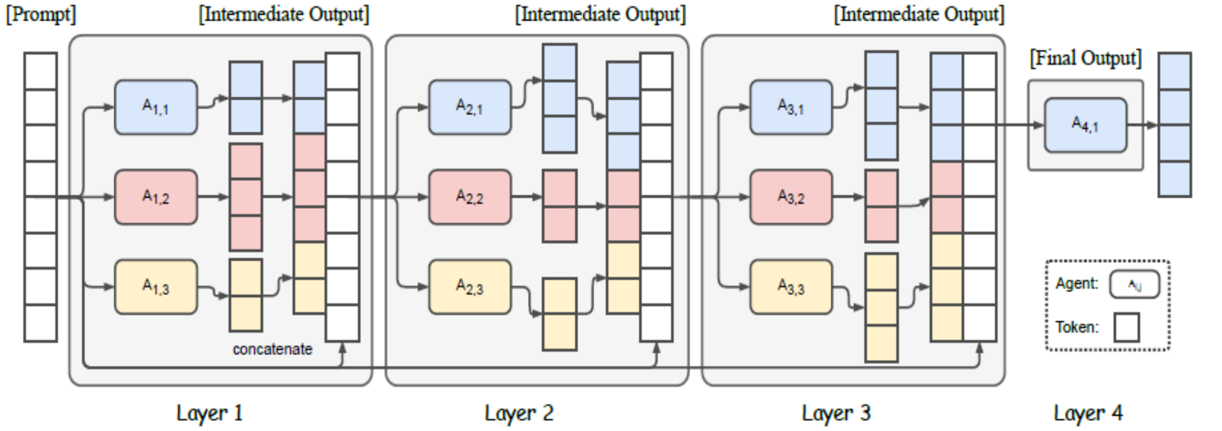


Figure 2: Illustration of the Mixture-of-Agents Structure. This example showcases 4 MoA layers with 3 agents in each layer. The agents here can share the same model.

- [Yang et al. (2024)] - **Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models**

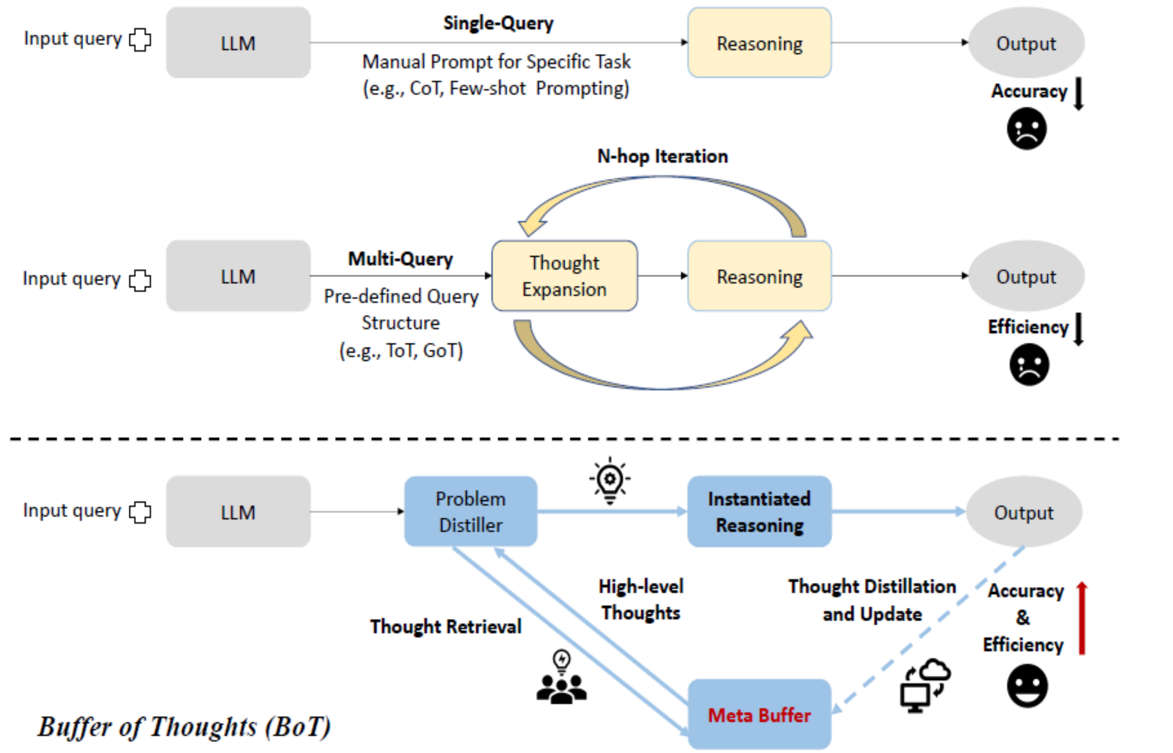
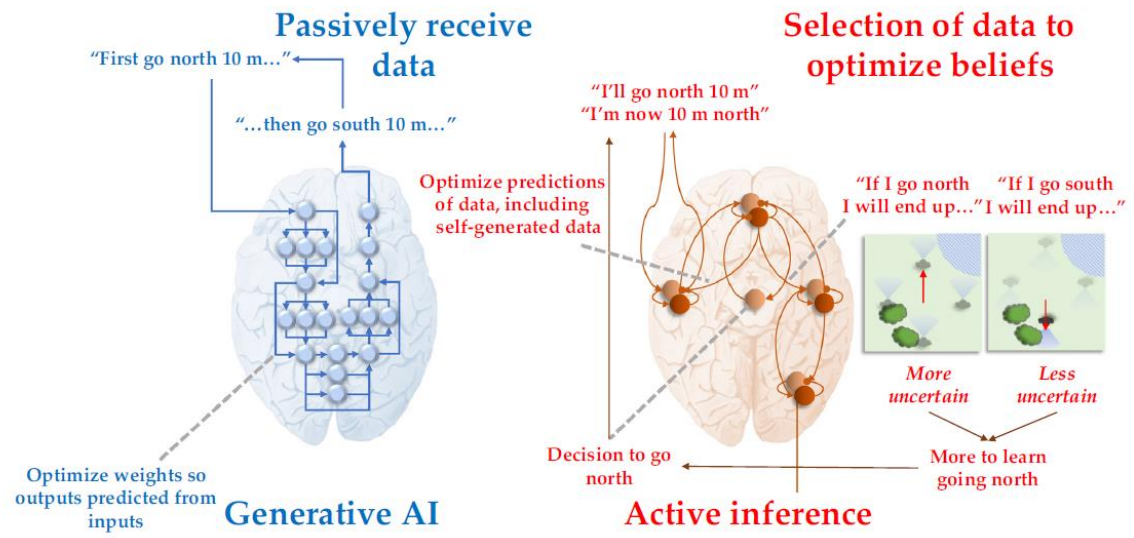


Figure 1: Comparison between single-query [8, 11], multi-query [14, 17], and (c) our BoT methods.

- [Pezzulo et al. (2024)] – **Generating meaning: active inference and the scope and limits of passive AI**


Passively receive data
"First go north 10 m…"
"…then go south 10 m…"
Optimize weights so outputs predicted from inputs
Generative AI
Selection of data to optimize beliefs
"I'll go north 10 m"
"I'm now 10 m north"
Optimize predictions of data, including self-generated data
"If I go north I will end up…"
"If I go south I will end up…"
More uncertain
Less uncertain
More to learn going north
Decision to go north
Active inference

Trends in Cognitive Sciences

Figure 2. How generative artificial intelligence (AI) and biological systems might learn generative models to solve the wayfinding task of Figure 1. (Left) Cartoon of the pretraining process for generative AI systems in which they are passively presented with (large quantities) of data. The weights of the network are then optimized such that their outputs are more probable given the inputs. State-of-the-art models often include subsequent fine-tuning in a (semi)supervised manner [88]; however, this still relies upon passive presentation of labeled data or self-generated outputs paired with rewards. (Right) By contrast, the generative models that underwrite active inference [148] involve reciprocal interactions with the world. This means that our current beliefs about the world can be used to select those data that have 'epistemic affordance' – in other words they are most useful to resolve our uncertainty about the data-generating process. In the process of learning what it means to go north or south, we may be more or less certain about the location we will end up in under each of these actions (shown here with a relatively high confidence of ending up in the southern position if going south, but more uncertainty in going north). By choosing to go north (and observing being 10 m north from our starting location), we are now in a better position to resolve our uncertainty and optimize our predictions. Beliefs about the causes of our data are an important part of this process of curiosity, exploration, or information seeking [80]. However, these beliefs may easily be neglected in the process of function approximation used in current generative AI systems, where all that matters is the desired output. The neuroanatomical diagrams in this figure are intended purely for illustrative purposes and are not to be taken seriously as anatomical hypotheses – which would distract from the focus of this paper on AI. However, process theories have been developed from active inference frameworks (e.g., [93,103,149]) to which we direct interested readers. Broadly, we might expect planning and policy selection to rely upon networks involving cortical and subcortical regions (e.g., cortico-basal-ganglia-thalamo-cortical loops) in which asymmetrical neuronal connectivity patterns between different cortical regions reflect communication between different hierarchical levels.

- [Rosenberg et al. (2024)] - **Collective Superintelligence: Amplifying Group IQ using Conversational Swarms** *(MAS}*

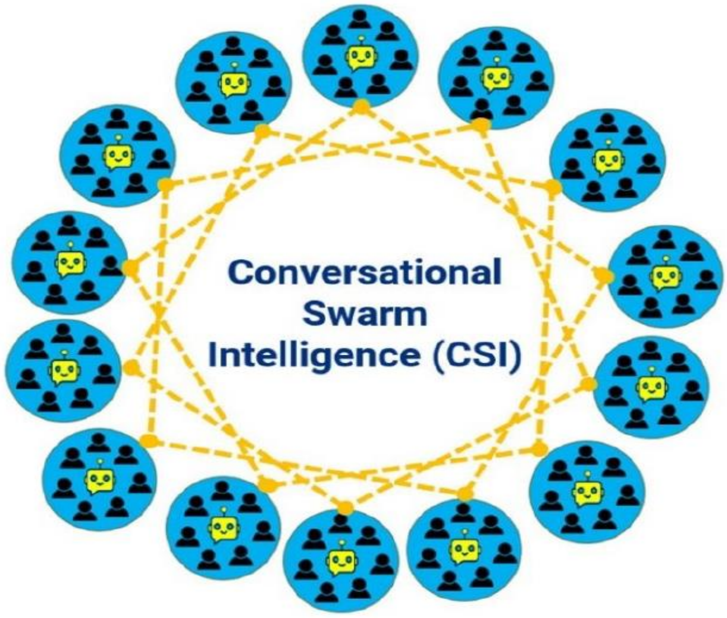


Fig. 1. Architecture for a Conversational Swarm Intelligence with AI agents assigned to each subgroup for passing and receiving conversastional content.

- [Lee et al. (2024)] – **MoAI: Mixture of All Intelligence for Large Language and Vision Models**

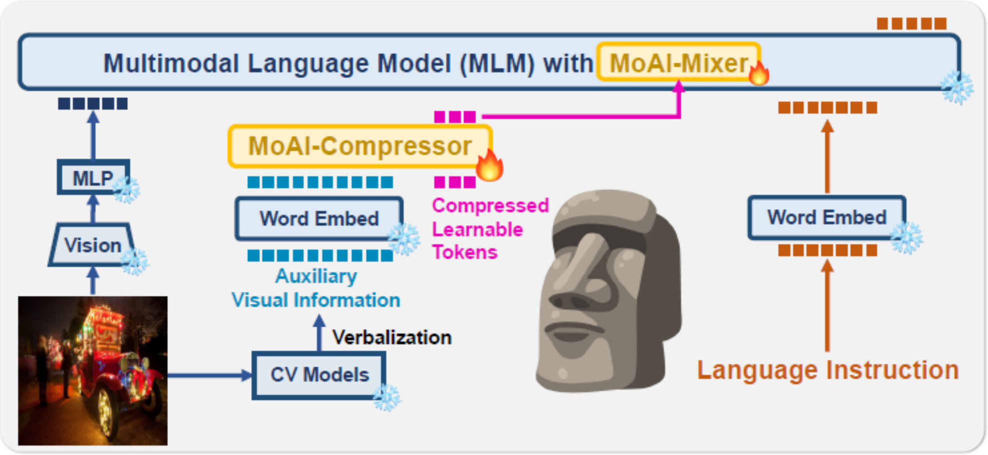


Fig. 3: Overview of 🗿 **MoAI** architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

(MAS}

- [Google DeepMind (2024a)] - **DiPaCo: Distributed Path Composition**

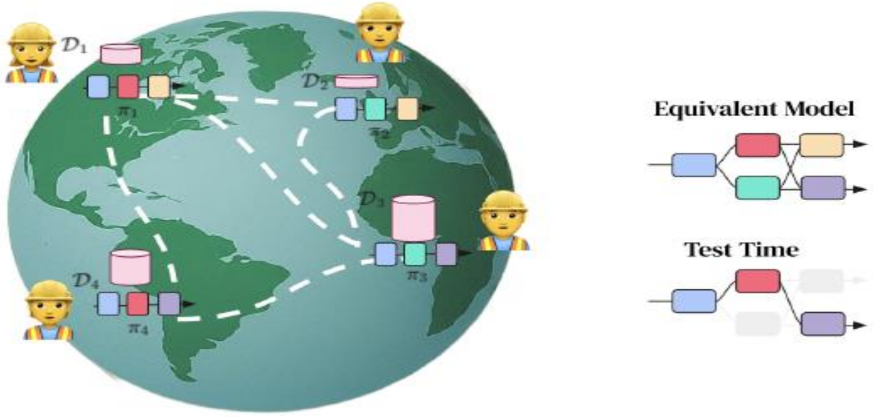
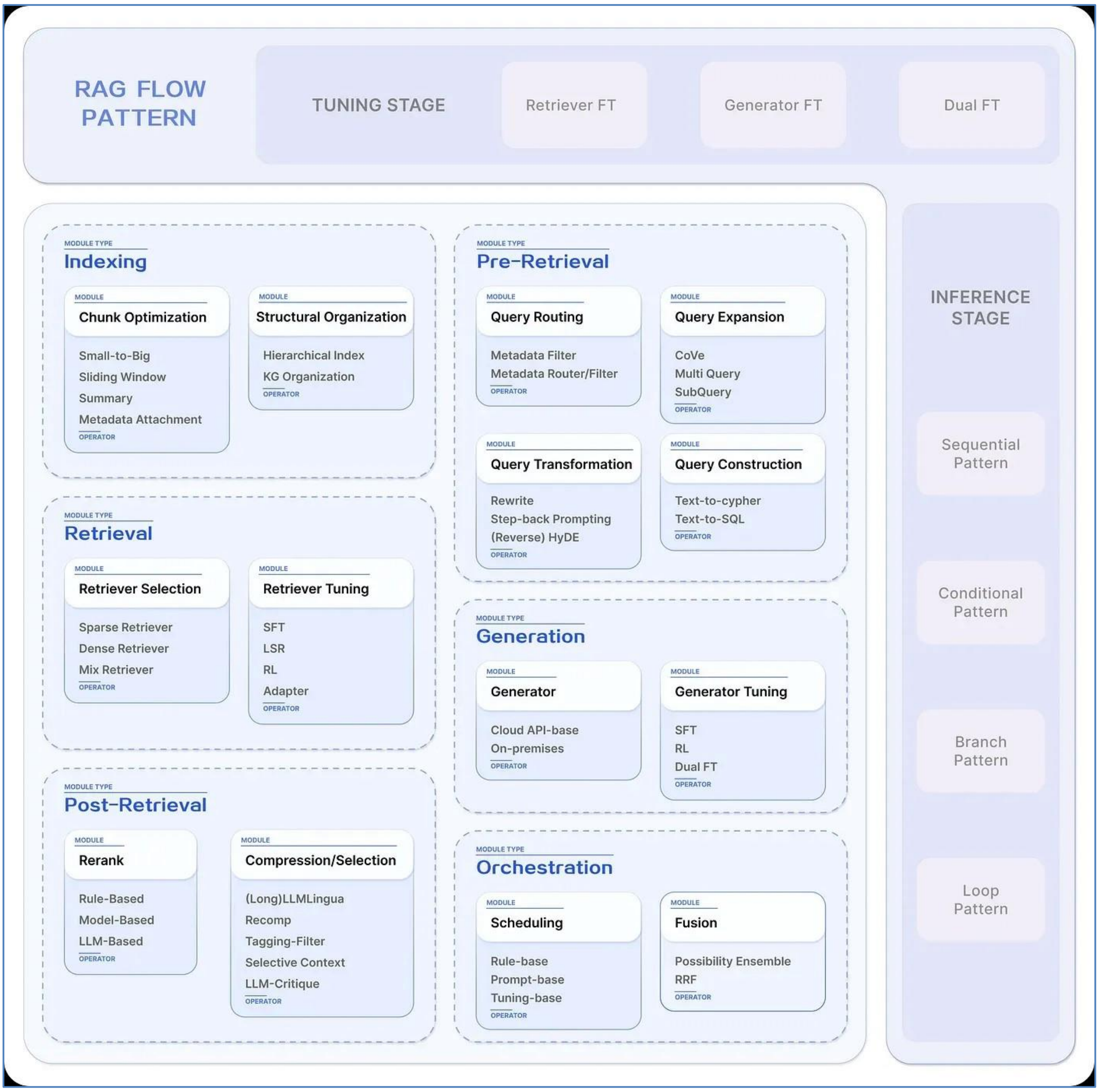
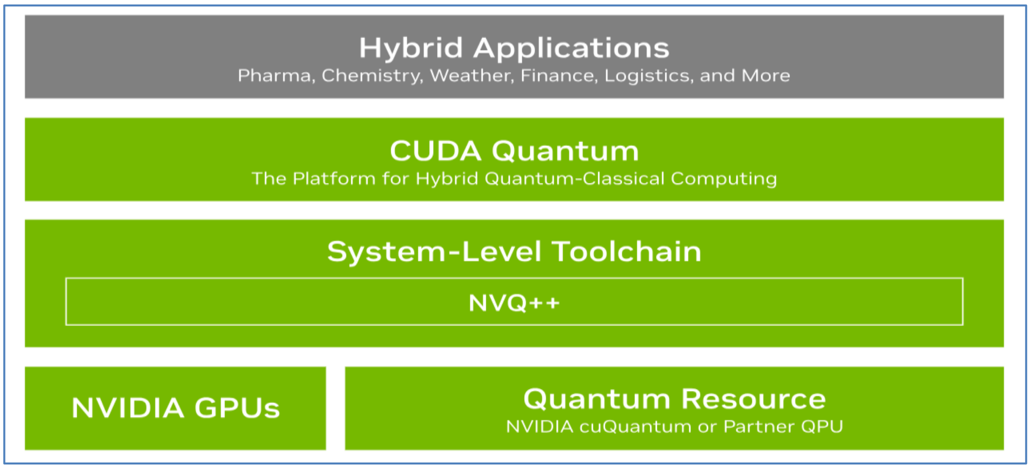


Figure 1 | **Long-term Goal:** Ultimately, we envision a modular network where different components, *paths* π_i, are optimized for different tasks, $\mathcal{D}_j$, each designed by different researchers. The paths, trained on any available hardware type, communicate infrequently across the world, exchanging useful information and enabling new forms of composition.

- [Zhaozhiming (2024)] - **Advanced RAG Retrieval Strategies: Flow and Modular**



- [NVIDIA (2024)] - **Hybrid and integrated systems and reference architecture for quantum-classical computing**


Hybrid Applications
Pharma, Chemistry, Weather, Finance, Logistics, and More
CUDA Quantum
The Platform for Hybrid Quantum-Classical Computing
System-Level Toolchain
NVQ++
NVIDIA GPUs
Quantum Resource
NVIDIA cuQuantum or Partner QPU

- [Sorensen et al. (2024)] – **A Roadmap to Pluralistic Alignment**
  (MAS}

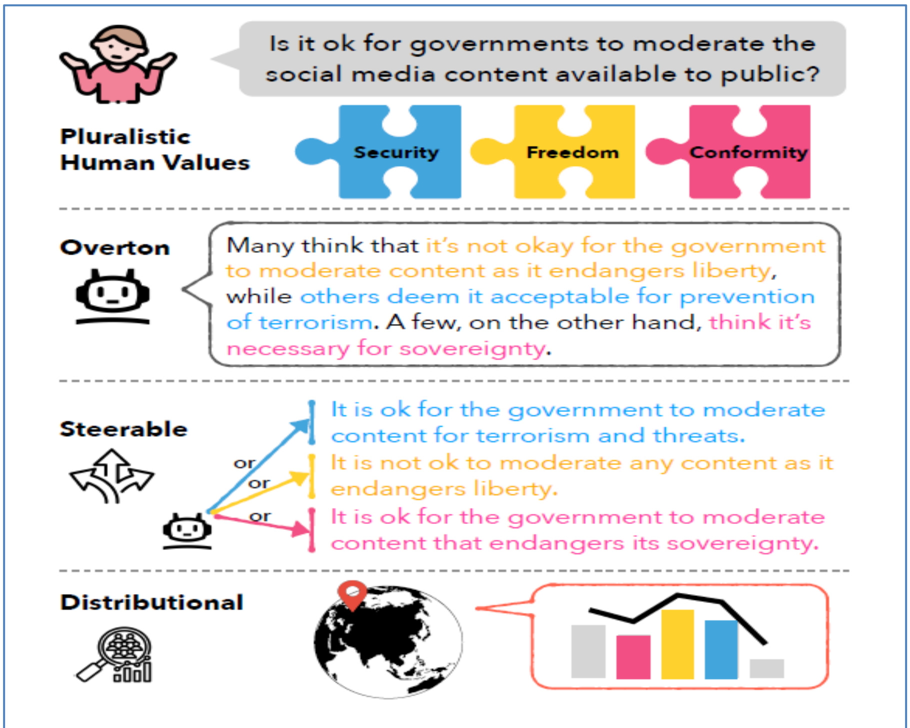

Is it ok for governments to moderate the social media content available to public?
Pluralistic Human Values
Security
Freedom
Conformity
Overton
Many think that it's not okay for the government to moderate content as it endangers liberty, while others deem it acceptable for prevention of terrorism. A few, on the other hand, think it's necessary for sovereignty.
Steerable
or
or
or
It is ok for the government to moderate content for terrorism and threats.
It is not ok to moderate any content as it endangers liberty.
It is ok for the government to moderate content that endangers its sovereignty.
Distributional

Figure 1. Three kinds of pluralism in models.

- [Masterman et al. (2024)] – **The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: a survey**

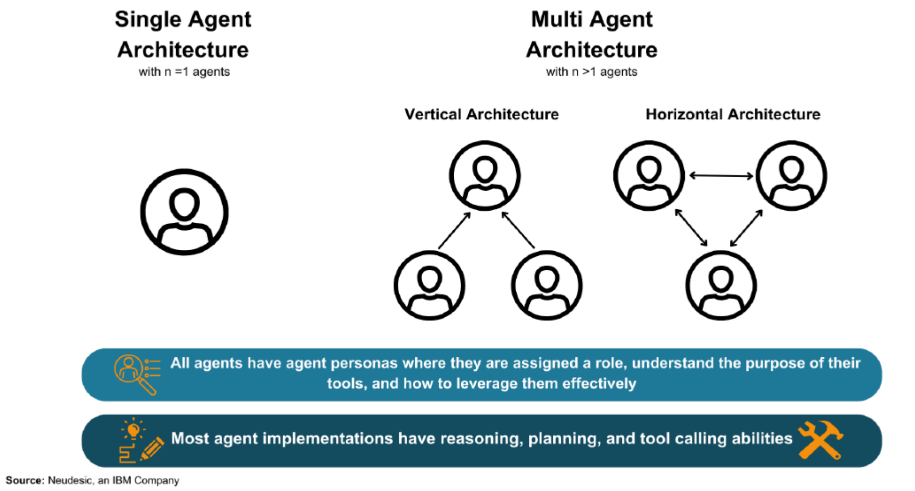
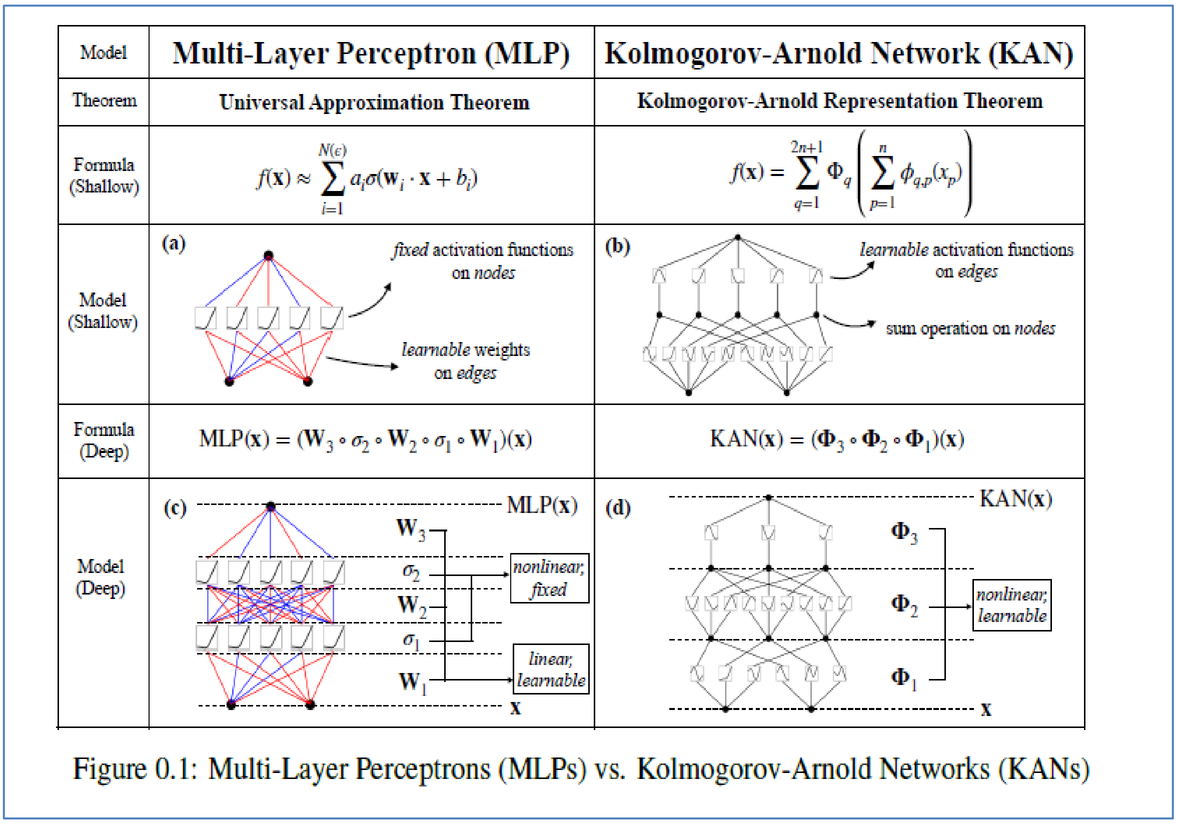

Single Agent Architecture
with n =1 agents
Multi Agent Architecture
with n >1 agents
Vertical Architecture
Horizontal Architecture
All agents have agent personas where they are assigned a role, understand the purpose of their tools, and how to leverage them effectively
Most agent implementations have reasoning, planning, and tool calling abilities
Source: Neudesic, an IBM Company

Figure 1: A visualization of single and multi-agent architectures with their underlying features and abilities

- [Liu et al. (2024)] – **KAN: Kolmogorov–Arnold Networks**



Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i\sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q\left(\sum_{p=1}^{n} \phi_{q,p}(x_p)\right)$
Model (Shallow)	(a) *fixed* activation functions on *nodes* *learnable* weights on *edges*	(b) *learnable* activation functions on *edges* sum operation on *nodes*
Formula (Deep)	$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\mathrm{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)	(c) MLP(x) $\mathbf{W}_3$ σ_2 $\mathbf{W}_2$ σ_1 $\mathbf{W}_1$ x *nonlinear, fixed* *linear, learnable*	(d) KAN(x) $\mathbf{\Phi}_3$ $\mathbf{\Phi}_2$ $\mathbf{\Phi}_1$ x *nonlinear, learnable*

Figure 0.1: Multi-Layer Perceptrons (MLPs) vs. Kolmogorov-Arnold Networks (KANs)

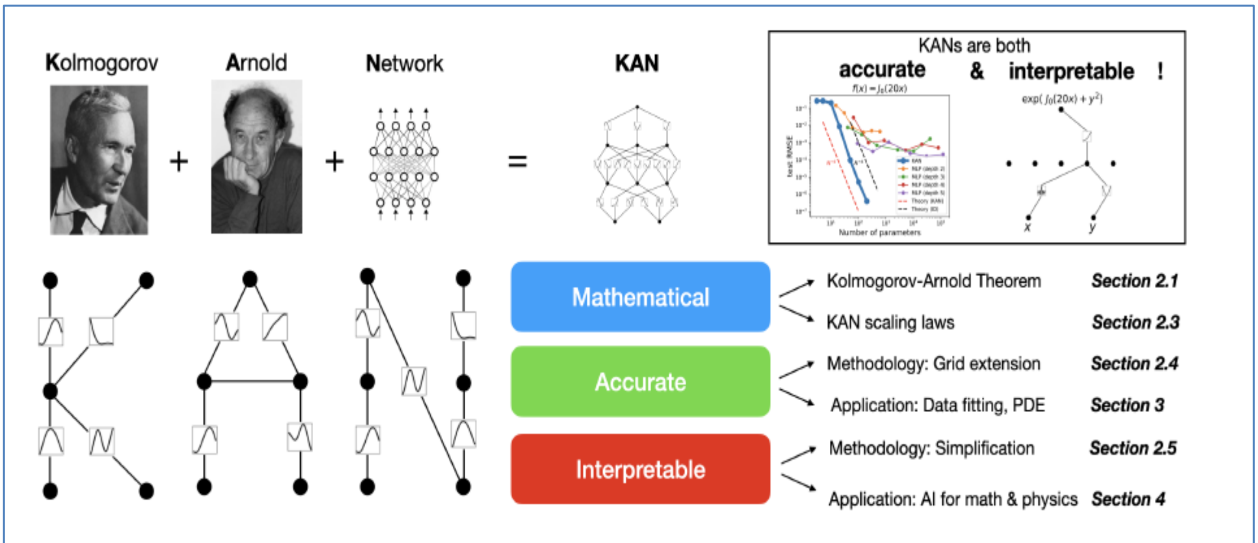
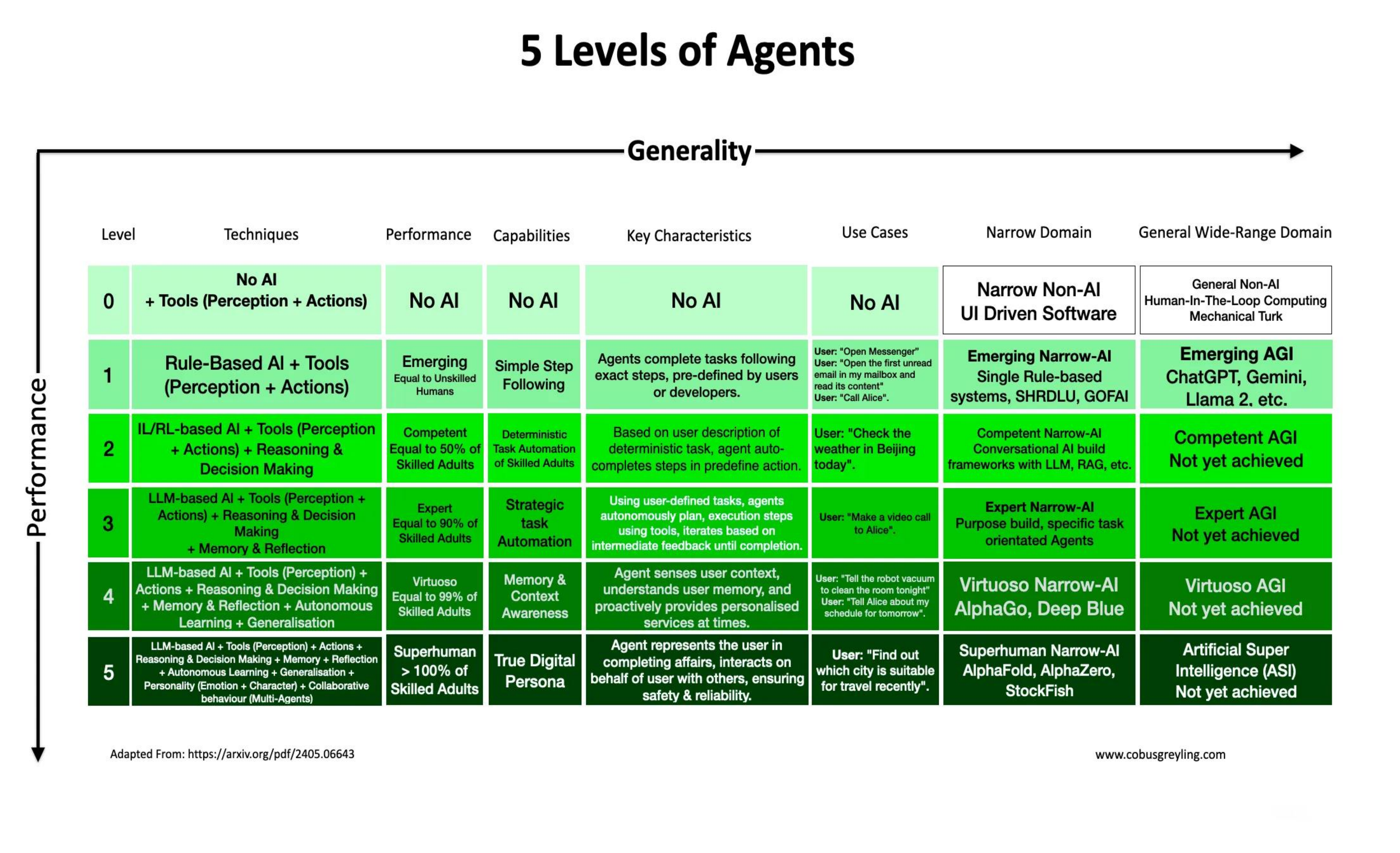

Kolmogorov
Arnold
Network
KAN
+
+
=
KANs are both
accurate & interpretable !
$f(x)=J_0(20x)$
test RMSE
Number of parameters
$\exp(J_0(20x)+y^2)$
x
y
Mathematical
Accurate
Interpretable
Kolmogorov-Arnold Theorem Section 2.1
KAN scaling laws Section 2.3
Methodology: Grid extension Section 2.4
Application: Data fitting, PDE Section 3
Methodology: Simplification Section 2.5
Application: AI for math & physics Section 4

Figure 2.1: Our proposed Kolmogorov-Arnold networks are in honor of two great late mathematicians, Andrey Kolmogorov and Vladimir Arnold. KANs are mathematically sound, accurate and interpretable.

- [Greyling (2024) and Huang (2024)] – **Levels Of AI Agents**

(See Table on next page)



5 Levels of Agents

Generality
Performance

Level	Techniques	Performance	Capabilities	Key Characteristics	Use Cases	Narrow Domain	General Wide-Range Domain
0	No AI + Tools (Perception + Actions)	No AI	No AI	No AI	No AI	Narrow Non-AI UI Driven Software	General Non-AI Human-In-The-Loop Computing Mechanical Turk
1	Rule-Based AI + Tools (Perception + Actions)	Emerging Equal to Unskilled Humans	Simple Step Following	Agents complete tasks following exact steps, pre-defined by users or developers.	User: "Open Messenger" User: "Open the first unread email in my mailbox and read its content" User: "Call Alice".	Emerging Narrow-AI Single Rule-based systems, SHRDLU, GOFAI	Emerging AGI ChatGPT, Gemini, Llama 2, etc.
2	IL/RL-based AI + Tools (Perception + Actions) + Reasoning & Decision Making	Competent Equal to 50% of Skilled Adults	Deterministic Task Automation of Skilled Adults	Based on user description of deterministic task, agent auto-completes steps in predefine action.	User: "Check the weather in Beijing today".	Competent Narrow-AI Conversational AI build frameworks with LLM, RAG, etc.	Competent AGI Not yet achieved
3	LLM-based AI + Tools (Perception + Actions) + Reasoning & Decision Making + Memory & Reflection	Expert Equal to 90% of Skilled Adults	Strategic task Automation	Using user-defined tasks, agents autonomously plan, execution steps using tools, iterates based on intermediate feedback until completion.	User: "Make a video call to Alice".	Expert Narrow-AI Purpose build, specific task orientated Agents	Expert AGI Not yet achieved
4	LLM-based AI + Tools (Perception) + Actions + Reasoning & Decision Making + Memory & Reflection + Autonomous Learning + Generalisation	Virtuoso Equal to 99% of Skilled Adults	Memory & Context Awareness	Agent senses user context, understands user memory, and proactively provides personalised services at times.	User: "Tell the robot vacuum to clean the room tonight" User: "Tell Alice about my schedule for tomorrow".	Virtuoso Narrow-AI AlphaGo, Deep Blue	Virtuoso AGI Not yet achieved
5	LLM-based AI + Tools (Perception) + Actions + Reasoning & Decision Making + Memory + Reflection + Autonomous Learning + Generalisation + Personality (Emotion + Character) + Collaborative behaviour (Multi-Agents)	Superhuman > 100% of Skilled Adults	True Digital Persona	Agent represents the user in completing affairs, interacts on behalf of user with others, ensuring safety & reliability.	User: "Find out which city is suitable for travel recently".	Superhuman Narrow-AI AlphaFold, AlphaZero, StockFish	Artificial Super Intelligence (ASI) Not yet achieved

Adapted From: https://arxiv.org/pdf/2405.06643

www.cobusgreyling.com

P. State of AI Research Roundups

Here we present 2024H2 Bi-Weekly Roundups (Latest research summaries in ML/DL, Robotics, CV, NLP and GenAI) from **[State of AI (2024)]** open public platform - Titles only (alphabetically) without any structuration, references, summaries and comments – to show the actual issues in this R&D area.

- A call for **embodied** AI
- A Comprehensive Evaluation of Histopathology Foundation Models for Ovarian Cancer Subtype Classification
- A Controlled Study on Long Context Extension and Generalization in LLMs
- A **Quantum** Leaky Integrate-and-Fire Spiking Neuron and Network
- A Single Transformer for Scalable Vision-Language Modeling
- A Monte Carlo Framework for Calibrated Uncertainty Estimation in Sequence Prediction
- Accumulator-Aware Post-Training Quantization
- ACEGEN: Reinforcement learning of generative chemical **agents** for drug discovery
- Action Contextualization: **Adaptive** Task Planning and Action Tuning using Large Language Models
- ActiveGS: Active Scene Reconstruction using Gaussian Splatting
- **Adaptive** Computation Modules: Granular Conditional Computation For Efficient **Inference**
- **Adaptive** Deployment of Untrusted LLMs Reduces Distributed Threats
- **Adaptive** Draft-Verification for Efficient Large Language Model Decoding
- **Adaptive** Caching for Faster Video Generation with Diffusion Transformers
- Additive-feature-attribution methods: a review on explainable artificial intelligence for fluid dynamics and heat transfer
- Addressing Uncertainty in LLMs to Enhance Reliability in Generative AI
- Adopting RAG for LLM-Aided Future Vehicle Design
- Advancing Fine-Grained Visual Understanding with Multi-Scale Alignment in **Multi-Modal** Models
- Adversarial Attacks on Large Language Models in Medicine
- Adversarially Robust Decision Transformer
- Aerial Assistive Payload Transportation Using Quadrotor UAVs with Nonsingular Fast Terminal SMC for Human Physical Interaction
- Affective Computing Has Changed: The Foundation Model Disruption
- **Agent** Instructs Large Language Models to be General Zero-Shot **Reasoners**
- **AGENT**-CQ: Automatic Generation and Evaluation of Clarifying Questions for Conversational Search with LLMs
- **AGENT**iGraph: An Interactive **Knowledge Graph** Platform for LLM-based Chatbots Utilizing Private Data
- Aguvis: Unified Pure Vision **Agents** for Autonomous GUI Interaction
- AI-accelerated discovery of high critical temperature superconductors
- AlphaRouter: **Quantum** Circuit Routing with Reinforcement Learning and Tree Search

- Altogether: Image Captioning via Re-aligning Alt-text
- An Introduction to **Quantum** Reinforcement Learning (QRL)
- Analog In-Memory Computing Attention Mechanism for Fast and Energy-Efficient Large Language Models
- ANOLE: An Open, Autoregressive, Native Large **Multimodal** Models for Interleaved Image-Text Generation
- AnyBipe: An End-to-End Framework for Training and Deploying Bipedal Robots Guided by Large Language Models
- AnyTaskTune: Advanced Domain-Specific Solutions through Task-Fine-Tuning
- APOLLO: SGD-like Memory, AdamW-level Performance
- Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models
- Approximation and generalization properties of the random projection classification method
- ARMADA: Attribute-Based **Multimodal** Data Augmentation
- As Generative Models Improve, People Adapt Their Prompts
- Assumption-Lean and Data-Adaptive Post-Prediction **Inference**
- Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models
- ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale
- Atari-GPT: Investigating the Capabilities of **Multimodal** Large Language Models as Low-Level Policies for Atari Games
- Attention Is All You Need But You Don't Need All Of It For **Inference** of Large Language Models
- Attention Prompting on Image for Large Vision-Language Models
- AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution
- AutoCodeRover: **Autonomous** Program Improvement
- AutoDefense: **Multi-Agent** LLM Defense against Jailbreak Attacks
- AutoGPT+P: **Affordance**-based Task Planning with Large Language Models
- Automated Ensemble **Multimodal** Machine Learning for Healthcare
- Automating the Search for Artificial Life with Foundation Models
- AutoScale: Automatic Prediction of Compute-optimal Data Composition for Training LLM
- AutoTurb: Using Large Language Models for Automatic Algebraic Model Discovery of Turbulence Closure
- Awaking the Slides: A Tuning-free and Knowledge-regulated AI Tutoring System via Language Model Coordination
- AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration
- BaichuanSEED: Sharing the Potential of ExtensivE Data Collection and Deduplication by Introducing a Competitive Large Language Model Baseline
- BALROG: Benchmarking **Agentic** LLM and VLM **Reasoning** On Games

- BAM! Just Like That: Simple and Efficient Parameter Upcycling for **Mixture of Experts**
- BEACON: Benchmark for Comprehensive RNA Tasks and Language Models
- BehaviorGPT: Smart **Agent** Simulation for Autonomous Driving with Next-Patch Prediction
- Bellman Diffusion: Generative Modeling as Learning a Linear Operator in the Distribution Space
- Best-of-N Jailbreaking
- Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools
- BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Representation Capabilities
- BIRD: A Trustworthy Bayesian **Inference** Framework for Large Language Models
- BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion
- BLAST: Block-Level **Adaptive** Structured Matrices for Efficient Deep Neural Network **Inference**
- Brain-like Functional Organization within Large Language Models
- Buckle Up: Robustifying LLMs at Every Customization Stage via Data Curation
- BUZZ: Beehive-structured Sparse KV Cache with Segmented Heavy Hitters for Efficient LLM **Inference**
- ByteCheckpoint: A Unified Checkpointing System for LLM Development
- Cambrian-1: A Fully Open, Vision-Centric Exploration of **Multimodal** LLMs
- CamemBERT 2.0: A Smarter French Language Model Aged to Perfection
- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers
- Can OpenSource beat ChatGPT? -- A Comparative Study of Large Language Models for Text-to-Code Generation
- Can Unconfident LLM Annotations Be Used for Confident Conclusions?
- Can Large Language Model **Agents** Simulate Human Trust Behavior?
- CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models
- Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models
- Castling-ViT: Compressing Self-Attention via Switching Towards Linear-Angular Attention at Vision Transformer Inference
- Causal Diffusion Transformers for Generative Modeling
- Causal Language Modeling Can Elicit Search and **Reasoning** Capabilities on Logic Puzzles
- Causal **Reasoning** and Large Language Models: Opening a New Frontier for Causality
- Cavia: Camera-controllable Multi-view Video Diffusion with View-Integrated Attention
- CDChat: A Large **Multimodal** Model for Remote Sensing Change Description
- Celtibero: Robust Layered Aggregation for Federated Learning
- Certified Robustness to Data Poisoning in Gradient-Based Training
- Chain of Code: **Reasoning** with a Language Model-Augmented Code Emulator
- ChartifyText: Automated Chart Generation from Data-Involved Texts via LLM
- ChatGarment: Garment Estimation, Generation and Editing via Large Language Models

- ChatGPT's Potential in Cryptography Misuse Detection: A Comparative Analysis with Static Analysis Tools
- CIBench: Evaluating Your LLMs with a Code Interpreter Plugin
- CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications
- CLEFT: Language-Image Contrastive Learning with Efficient Large Language Model and Prompt Fine-Tuning
- CodeGRAG: Bridging the Gap between Natural Language and Programming Language via Graphical Retrieval Augmented Generation
- CodeLutra: Boosting LLM Code Generation via Preference-Guided Refinement
- CodeV: Empowering LLMs for Verilog Generation through Multi-Level Summarization
- Cognitive Kernel: An Open-source **Agent** System towards Generalist Autopilots
- CogVLM2: Visual Language Models for Image and Video Understanding
- Combining Large Language Models with Tutoring System Intelligence: A Case Study in Caregiver Homework Support
- COMET: Benchmark for Comprehensive Biological Multi-omics Evaluation Tasks and Language Models
- Comprehensive benchmarking of large language models for RNA secondary structure prediction
- Compress then Serve: Serving Thousands of LoRA Adapters with Little Overhead
- Compressed Chain of Thought: Efficient **Reasoning** Through Dense Representations
- Compute-Constrained Data Selection
- Conditional Forecasting of Margin Calls using Dynamic Graph Neural Networks
- Constraining Generative Models for Engineering Design with Negative Data
- **Continual Memorization** of Factoids in Large Language Models
- Controlling Counterfactual Harm in Decision Support Systems Based on Prediction Sets
- Converging Paradigms: The **Synergy of Symbolic and Connectionist AI** in LLM-Empowered Autonomous **Agents**
- ConvMixFormer- A Resource-efficient Convolution Mixer for Transformer-based Dynamic Hand Gesture Recognition
- CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation
- Corki: Enabling Real-time **Embodied** AI Robots via Algorithm-Architecture Co-Design
- CRAB: Cross-environment **Agent** Benchmark for **Multimodal** Language Model Agents
- Creative Beam Search: LLM-as-a-Judge For Improving Response Generation
- CREW: Facilitating Human-AI Teaming Research
- CRITERIA: a New Benchmarking Paradigm for Evaluating Trajectory Prediction Models for Autonomous Driving
- Croissant: A Metadata Format for ML-Ready Datasets

- CrowdMoGen: Zero-Shot Text-Driven Collective Motion Generation
- CTIBench: A Benchmark for Evaluating LLMs in Cyber Threat Intelligence
- DART: Denoising Autoregressive Transformer for Scalable Text-to-Image Generation
- Data Interpreter: An LLM **Agent** For Data Science
- DataGpt-SQL-7B: An Open-Source Language Model for Text-to-SQL
- DeeR-VLA: Dynamic **Inference** of **Multimodal** Large Language Models for Efficient Robot Execution
- Defending Our Privacy With Backdoors
- Demo: SGCode: A Flexible **Prompt-Optimizing** System for Secure Generation of Code
- Denoising diffusion models for high-resolution microscopy image restoration
- Derivational Morphology Reveals Analogical Generalization in Large Language Models
- Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding
- DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning
- Differential Privacy Regularization: Protecting Training Data Through Loss Function Regularization
- DiffH2O: Diffusion-Based Synthesis of Hand-Object Interactions from Textual Descriptions
- Diffusing States and Matching Scores: A New Framework for Imitation Learning
- Diffusion Attribution Score: Evaluating Training Data Influence in Diffusion Model
- Diffusion Language Models Are Versatile Protein Learners
- Diffusion Self-Distillation for Zero-Shot Customized Image Generation
- Diffusion Transformer Policy
- DINO Pre-training for Vision-based End-to-end Autonomous Driving
- Disability data futures: Achievable imaginaries for AI and disability data justice
- DISCO: Efficient Diffusion Solver for Large-Scale Combinatorial Optimization Problems
- **Disentangling Memory** and **Reasoning** Ability in Large Language Models
- Disrupting Test Development with AI Assistants
- Distilling System 2 into System 1
- Diverse Score Distillation
- Dockformer: A transformer-based molecular docking paradigm for large-scale virtual screening
- Does CLIP Know My Face?
- DoRA: Weight-Decomposed Low-Rank **Adaptation**
- DPI-TTS: Directional Patch Interaction for Fast-Converging and Style Temporal Modeling in Text-to-Speech
- DreamGarden: A Designer Assistant for Growing Games from a Single Prompt
- DreamHead: Learning Spatial-Temporal Correspondence via Hierarchical Diffusion for Audio-driven Talking Head Synthesis
- DreamHOI: Subject-Driven Generation of 3D Human-Object Interactions with Diffusion Priors
- DreamVideo: High-Fidelity Image-to-Video Generation with Image Retention and Text Guidance

- DreamVTON: Customizing 3D Virtual Try-on with Personalized Diffusion Models
- DreamWaltz-G: Expressive 3D Gaussian Avatars from Skeleton-Guided 2D Diffusion
- DressRecon: Freeform 4D Human Reconstruction from Monocular Video
- DroidSpeak: KV Cache Sharing for Efficient **Multi-LLM** Serving
- DuoAttention: Efficient Long-Context LLM **Inference** with Retrieval and Streaming Heads
- DuQuant: Distributing Outliers via Dual Transformation Makes Stronger Quantized LLMs
- Dynamic Memory Compression: Retrofitting LLMs for Accelerated **Inference**
- Eagle: Exploring The Design Space for **Multimodal** LLMs with **Mixture of Encoders**
- EasyControl: Transfer ControlNet to Video Diffusion for Controllable Generation and Interpolation
- Edify Image: High-Quality Image Generation with Pixel Space Laplacian Diffusion Models
- Efficient Feature Interactions with Transformers: Improving User Spending Propensity Predictions in Gaming
- Efficient **Multi-modal** Large Language Models via Visual Token Grouping
- Efficient End-to-End 6-Dof Grasp Detection Framework for **Edge Devices** with Hierarchical Heatmaps and Feature Propagation
- **Embodied Agent** Interface: Benchmarking LLMs for Embodied Decision Making
- **Emergence of Hidden Capabilities**: Exploring Learning Dynamics in Concept Space
- EMMA: End-to-End **Multimodal** Model for Autonomous Driving
- Empowering Clients: Transformation of Design Processes Due to Generative AI
- Empowering Robot Path Planning with Large Language Models: osmAG **Map Topology** & Hierarchy Comprehension with LLMs
- Encapsulating Knowledge in One Prompt
- End-to-End Navigation with Vision Language Models: Transforming **Spatial Reasoning** into Question-Answering
- EnergyDiff: Universal Time-Series Energy Data Generation using Diffusion Models
- Enhance **Reasoning** by Learning from Mistakes: Peer-Review Knowledge Distillation from **Multiple** Large Language Models
- Enhancing AI Accessibility in Veterinary Medicine: Linking Classifiers and Electronic Health Records
- Enhancing LLM **Reasoning** via Critique Models with Test-Time and Training-Time Supervision
- Enhancing Training Efficiency Using Packing with Flash Attention
- Entendre, a Social Bot Detection Tool for Niche, Fringe, and Extreme Social Media
- Entropic Distribution Matching in Supervised Fine-tuning of LLMs: Less Overfitting and Better Diversity
- EoRA: Training-free Compensation for Compressed LLM with Eigenspace Low-Rank Approximation
- Equity in the Use of ChatGPT for the Classroom: A Comparison of the Accuracy and Precision of ChatGPT 3.5 vs. ChatGPT4 with Respect to Statistics and Data Science Exams
- Espresso: Robust Concept Filtering in Text-to-Image Models
- EuroLLM: Multilingual Language Models for Europe

- Evolutionary Greedy Algorithm for Optimal Sensor Placement Problem in Urban Sewage Surveillance
- **Evolutionary** Reinforcement Learning via **Cooperative Coevolution**
- Examination of Code generated by Large Language Models
- Examining Imbalance Effects on Performance and Demographic Fairness of Clinical Language Models
- EXAONE 3.0 7.8B Instruction Tuned Language Model
- Executing Arithmetic: Fine-Tuning Large Language Models as Turing Machines
- Exploiting Student Parallelism for Low-latency GPU **Inference** of BERT-like Models in Online Services
- ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies
- Exploring AI Text Generation, Retrieval-Augmented Generation, and Detection Technologies: a Comprehensive Overview
- EyeCLIP: A visual-language foundation model for **multi-modal** ophthalmic image analysis
- FabGPT: An Efficient Large **Multimodal** Model for Complex Wafer Defect Knowledge Queries
- Fast Multipole Attention: A Divide-and-Conquer Attention Mechanism for Long Sequences
- FastCLIP: A Suite of Optimization Techniques to Accelerate CLIP Training with Limited Resources
- FasterCache: Training-Free Video Diffusion Model Acceleration with High Quality
- FASTNav: Fine-tuned **Adaptive** Small-language-models Trained for Multi-point Robot Navigation
- FBI-LLM: Scaling Up Fully Binarized LLMs from Scratch via Autoregressive Distillation
- FineZip : Pushing the Limits of Large Language Models for Practical Lossless Text Compression
- FiTv2: Scalable and Improved Flexible Vision Transformer for Diffusion Model
- Flash Communication: Reducing Tensor Parallelization Bottleneck for Fast Large Language Model **Inference**
- Flextron: Many-in-One Flexible Large Language Model
- Flowing from Words to Pixels: A Framework for **Cross-Modality** Evolution
- Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens
- FocusLLM: Scaling LLM's Context by Parallel Decoding
- ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities
- Foundation Models for the Electric Power Grid
- Foundational Large Language Models for Materials Research
- Free-Mask: A Novel Paradigm of Integration Between the Segmentation Diffusion Model and Image Editing to Improve Segmentation Ability
- Frequency Adaptive Normalization For Non-stationary Time Series Forecasting
- From Decoding to Meta-Generation: **Inference**-time Algorithms for Large Language Models
- From Feature Importance to Natural Language Explanations Using LLMs with RAG
- From Prompt Engineering to Prompt Craft
- FrontierMath: A Benchmark for Evaluating Advanced Mathematical **Reasoning** in AI
- FullStack Bench: Evaluating LLMs as Full Stack Coders

- Gaussian is All You Need: A Unified Framework for Solving Inverse Problems via Diffusion Posterior Sampling
- GaussianAnything: Interactive Point Cloud Latent Diffusion for 3D Generation
- GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative **Inference** of LLM
- Gemma 2: Improving Open Language Models at a Practical Size
- GenCode: A Generic Data Augmentation Framework for Boosting Deep Learning-Based Code Understanding
- Generalization of Graph Neural Networks is Robust to Model Mismatch
- Generalized Robot Learning Framework
- Generative AI in Medicine
- Generative **Semantic** Communication: Architectures, Technologies, and Applications
- Generative Verifiers: Reward Modeling as Next-Token Prediction
- GenRec: Unifying Video Generation and Recognition with Diffusion Models
- GenSim2: Scaling Robot Data Generation with **Multi-modal and Reasoning** LLMs
- GeoCode-GPT: A Large Language Model for Geospatial Code Generation Tasks
- Geometric Representation Condition Improves Equivariant Molecule Generation
- Geometry Informed Tokenization of Molecules for Language Model Generation
- GIS Copilot: Towards an Autonomous GIS **Agent** for Spatial Analysis
- GLOV: Guided Large Language Models as Implicit Optimizers for Vision Language Models
- Goetterfunke: **Creativity** in Machinae Sapiens. About the Qualitative Shift in Generative AI with a Focus on Text-To-Image
- GRAB: A Challenging GRaph Analysis Benchmark for Large Multimodal Models
- Grounded Answers for **Multi-agent** Decision-making Problem through Generative **World Model**
- Grounding Large Language Models In **Embodied** Environment With Imperfect **World Models**
- Guiding Through Complexity: What Makes Good Supervision for Hard **Reasoning** Tasks?
- Gymnasium: A Standard Interface for Reinforcement Learning Environments
- Hands-On Tutorial: Labeling with LLM and Human-in-the-Loop
- Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey
- Harmonic LLMs are Trustworthy
- Harnessing Diversity for Important Data Selection in Pretraining Large Language Models
- Health AI Developer Foundations
- Health-LLM: Personalized Retrieval-Augmented Disease Prediction System
- Helping LLMs Improve Code Generation Using Feedback from Testing and Static Analysis
- Hermes: Memory-Efficient Pipeline **Inference** for Large Models on **Edge Devices**
- HiFi-CS: Towards Open Vocabulary Visual Grounding For Robotic Grasping Using Vision-Language Models
- How Diffusion Models Learn to Factorize and Compose

- Human Perception of LLM-generated Text Content in Social Media Environments
- Hunyuan-Large: An Open-Source MoE Model with 52 Billion Activated Parameters by Tencent
- HydraViT: Stacking Heads for a Scalable ViT
- Hyp2Nav: Hyperbolic Planning and Curiosity for Crowd Navigation
- Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation
- IDs for AI Systems
- Imagen 3
- Imperfect Vision Encoders: Efficient and Robust Tuning for Vision-Language Models
- Imposter.AI: Adversarial Attacks with Hidden Intentions towards Aligned Large Language Models
- Improving Alignment and Robustness with Circuit Breakers
- Improving Antibody Design with Force-Guided Sampling in Diffusion Models
- Improving Long-Text Alignment for Text-to-Image Diffusion Models
- Improving Ontology Requirements Engineering with OntoChat and Participatory Prompting
- Improving Pretraining Data Using Perplexity Correlations
- IncidentResponseGPT: Generating Traffic Incident Response Plans with Generative Artificial Intelligence
- Inclusive Design of AI's Explanations: Just for Those Previously Left Out, or for Everyone?
- **Inference** Scaling Laws: An Empirical Analysis of Compute-Optimal **Inference** for Problem-Solving with Language Models
- Instruction-Driven Game Engines on Large Language Models
- Instruct-SkillMix: A Powerful Pipeline for LLM Instruction Tuning
- InterFusion: Text-Driven Generation of 3D Human-Object Interaction
- Interpretable Graph Neural Networks for Heterogeneous Tabular Data
- Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations
- Introducing the Large Medical Model: State of the art healthcare cost and risk prediction with transformers trained on patient event sequences
- InvDesFlow: An AI search engine to explore possible high-temperature superconductors
- Jack of All Trades, Master of Some, a Multi-Purpose Transformer **Agent**
- Jamba-1.5: Hybrid Transformer-Mamba Models at Scale
- JEAN: Joint Expression and Audio-guided NeRF-based Talking Face Generation
- JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation
- Kernel Memory Networks: A Unifying Framework for **Memory Modeling**
- Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling
- **Knowledge Mechanisms** in Large Language Models: A Survey and Perspective
- KNOWNET: Guided Health Information Seeking from LLMs via **Knowledge Graph Integration**
- KOSMOS-2.5: A **Multimodal** Literate Model
- Large Language Model Safety: A Holistic Survey

- Large Language Models and Games: A Survey and Roadmap
- Large Language Models for Code: Security Hardening and Adversarial Testing
- Large Language Monkeys: Scaling **Inference** Compute with Repeated Sampling
- LayerKV: Optimizing Large Language Model Serving with Layer-wise KV Cache Management
- LCFO: Long Context and Long Form Output Dataset and Benchmarking
- Learning Novel Skills from Language-Generated Demonstrations
- Learning To Help: Training Models to Assist Legacy Devices
- Learning to Price Homogeneous Data
- Learning with Less: Knowledge Distillation from Large Language Models via Unlabeled Data
- Let's Get to the Point: LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts
- Leveraging Chemistry Foundation Models to Facilitate Structure Focused Retrieval Augmented Generation in **Multi-Agent** Workflows for Catalyst and Materials Design
- Leveraging **Mixture of Experts** for Improved Speech Deepfake Detection
- Likelihood as a Performance Gauge for Retrieval-Augmented Generation
- Liquid: Language Models are Scalable **Multi-modal** Generators
- LlamaFusion: Adapting Pretrained Language Models for **Multimodal** Generation
- LLaMA-Mesh: Unifying 3D Mesh Generation with Language Models
- LLaMA-Omni: Seamless Speech Interaction with Large Language Models
- LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages
- LLAssist: Simple Tools for Automating Literature Review Using Large Language Models
- LLaVA-MoD: Making LLaVA Tiny via **MoE** Knowledge Distillation
- LLM Echo Chamber: personalized and automated disinformation
- LLM Hallucinations in Practical Code Generation: Phenomena, Mechanism, and Mitigation
- LLM Pruning and Distillation in Practice: The Minitron Approach
- LLM2CLIP: Powerful Language Model Unlocks Richer Visual Representation
- LLM4DSR: Leveraing Large Language Model for Denoising Sequential Recommendation
- LLM-Craft: Robotic Crafting of Elasto-Plastic Objects with Large Language Models
- LLMmap: Fingerprinting For Large Language Models
- LLMPhy: Complex Physical **Reasoning** Using Large Language Models and **World Models**
- LLMs can realize combinatorial **creativity**: generating creative ideas via LLMs for scientific research
- LLMStinger: Jailbreaking LLMs using RL fine-tuned LLMs
- Local deployment of large-scale music AI models on commodity hardware
- Logic Query of Thoughts: Guiding Large Language Models to Answer Complex Logic Queries with **Knowledge Graphs**
- Long-form music generation with latent diffusion

- Long-Form Text-to-Music Generation with Adaptive Prompts: A Case of Study in Tabletop Role-Playing Games Soundtracks
- LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering
- LongVILA: Scaling Long-Context Visual Language Models for Long Videos
- Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps
- LoTLIP: Improving Language-Image Pre-training for Long Text Understanding
- Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens
- Low-Cost Language Models: Survey and Performance Evaluation on Python Code Generation
- LUMIA: Linear probing for Unimodal and **MultiModal** Membership **Inference** Attacks leveraging internal LLM states
- MAGIC: Generating **Self-Correction** Guideline for In-Context Text-to-SQL
- MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation with Speculative Decoding
- MagicPIG: LSH Sampling for Efficient LLM Generation
- MALMM: **Multi-Agent** Large Language Models for Zero-Shot Robotics Manipulation
- Mamba or Transformer for Time Series Forecasting? **Mixture of Universals** (MoU) Is All You Need
- Manifold-Constrained Nucleus-Level Denoising Diffusion Model for Structure-Based Drug Design
- **Mapping** the Unseen: Unified Promptable Panoptic Mapping with Dynamic Labeling using Foundation Models
- Marconi: Prefix Caching for the Era of Hybrid LLMs
- MarkLLM: An Open-Source Toolkit for LLM Watermarking
- Matryoshka **Multimodal** Models
- Measuring Bullshit in the Language Games played by ChatGPT
- Med-Bot: An AI-Powered Assistant to Provide Accurate and Reliable Medical Information
- MEGA-Bench: Scaling **Multimodal** Evaluation to over 500 Real-World Tasks
- Mélange: Cost Efficient Large Language Model Serving by Exploiting GPU Heterogeneity
- Melody Is All You Need For Music Generation
- MetaGPT: Meta Programming for A **Multi-Agent** Collaborative Framework
- MetaUrban: A Simulation Platform for **Embodied** AI in Urban Spaces
- MetricGold: Leveraging Text-To-Image Latent Diffusion Models for Metric Depth Estimation
- MHRC: Closed-loop Decentralized Multi-Heterogeneous Robot **Collaboration** with Large Language Models
- Mindalogue: LLM -- Powered Nonlinear Interaction for Effective Learning and Task Exploration
- MINERS: Multilingual Language Models as **Semantic** Retrievers

- MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention
- Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming
- MINT-1T: Scaling Open-Source **Multimodal** Data by 10x: A Multimodal Dataset with One Trillion Tokens
- Mix Data or Merge Models? Optimizing for Diverse Multi-Task Learning
- MLLM-LLaVA-FL: **Multimodal** Large Language Model Assisted Federated Learning
- MM1.5: Methods, Analysis & Insights from **Multimodal** LLM Fine-tuning
- MM-Ego: Towards Building Egocentric **Multimodal** LLMs
- MMEvol: Empowering **Multimodal** Large Language Models with Evol-Instruct
- **MoE**-Infinity: Offloading-Efficient MoE Model Serving
- MoFO: Momentum-Filtered Optimizer for Mitigating Forgetting in LLM Fine-Tuning
- Molecular Topological Profile (MOLTOP) -- Simple and Strong Baseline for Molecular Graph Classification
- Molmo and PixMo: Open Weights and Open Data for State-of-the-Art **Multimodal** Models
- Monet: **Mixture of Monosemantic Experts** for Transformers
- MonoFormer: One Transformer for Both Diffusion and Autoregression
- Mono-InternVL: Pushing the Boundaries of Monolithic **Multimodal** Large Language Models with Endogenous Visual Pre-training
- MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequence
- Multi-GraspLLM: A **Multimodal** LLM for Multi-Hand **Semantic** Guided Grasp Generation
- **Multimodal** Latent Language Modeling with Next-Token Diffusion
- Multi-Object Hallucination in Vision-Language Models
- Multi-view biomedical foundation models for molecule-target and property prediction
- MUVO: A **Multimodal World Model** with Spatial Representations for Autonomous Driving
- Natural Language Programming in Medicine: Administering Evidence Based Clinical Workflows with Autonomous **Agents** Powered by Generative Large Language Models
- Natural Language to Verilog: Design of a Recurrent Spiking Neural Network using Large Language Models and ChatGPT
- Navigating Extremes: Dynamic Sparsity in Large Output Space
- Navigating the Risks: A Survey of Security, Privacy, and Ethics Threats in LLM-Based **Agents**
- Neuro-Vision to Language: Enhancing Brain Recording-based Visual Reconstruction and Language Interaction
- Non-autoregressive Generative Models for Reranking Recommendation
- Non-discrimination Criteria for Generative Language Models
- Not All Layers of LLMs Are Necessary During **Inference**
- NT-LLM: A Novel Node Tokenizer for Integrating Graph Structure into Large Language Models
- N-Version Assessment and Enhancement of Generative AI

- NYU CTF Dataset: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security
- Obfuscated Activations Bypass LLM Latent-Space Defenses
- Occlusion-Aware Seamless Segmentation
- OKAMI: Teaching Humanoid Robots Manipulation Skills through Single Video Imitation
- Olympus: A Universal Task Router for Computer Vision Tasks
- OmniBench: Towards The Future of Universal Omni-Language Models
- OmniQuery: Contextually Augmenting Captured **Multimodal Memory** to Enable Personal Question Answering
- On Speeding Up Language Model Evaluation
- One-Step Diffusion Policy: Fast Visuomotor Policies via Diffusion Distillation
- Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models
- OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs
- Open-MAGVIT2: An Open-Source Project Toward Democratizing Auto-regressive Visual Generation
- Open-Source Conversational AI with SpeechBrain 1.0
- Order of Magnitude Speedups for LLM Membership **Inference**
- OS-ATLAS: A Foundation Action Model for Generalist GUI Agents
- P/D-Serve: Serving Disaggregated Large Language Model at Scale
- Pangea: A Fully Open Multilingual **Multimodal** LLM for 39 Languages
- ParaGAN: A Scalable Distributed Training Framework for Generative Adversarial Networks
- Pareto Data Framework: Steps Towards Resource-Efficient Decision Making Using Minimum Viable Data (MVD)
- Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits
- PEER: Expertizing Domain-Specific Tasks with a **Multi-Agent** Framework and Tuning Methods
- Persistent Pre-Training Poisoning of LLMs
- PICL: Physics Informed Contrastive Learning for Partial Differential Equations
- PIM-AI: A Novel Architecture for High-Efficiency LLM **Inference**
- POEM: Interactive Prompt Optimization for Enhancing Multimodal Reasoning of Large Language Models
- PokeFlex: Towards a Real-World Dataset of Deformable Objects for Robotic Manipulation
- Polaris: Open-ended Interactive Robotic Manipulation via Syn2Real Visual Grounding and Large Language Models
- Polymetis: Large Language Modeling for Multiple Material Domains
- PortLLM: Personalizing Evolving Large Language Models with Training-Free and Portable Model Patches
- Potential Based Diffusion Motion Planning

- PPT: Pre-Training with Pseudo-Labeled Trajectories for Motion Forecasting
- Preble: Efficient Distributed Prompt Scheduling for LLM Serving
- PrefixQuant: Static Quantization Beats Dynamic through Prefixed Outliers in LLMs
- Prioritized Generative Replay
- Programming Every Example: Lifting Pre-training Data Quality like Experts at Scale
- Promptable Closed-loop Traffic Simulation
- Prompt-Based Segmentation at Multiple Resolutions and Lighting Conditions using Segment Anything Model 2
- Prompting with Phonemes: Enhancing LLM Multilinguality for non-Latin Script Languages
- Proposer-Agent-Evaluator(PAE): Autonomous Skill Discovery For Foundation Model Internet **Agents**
- ProtoSAM - One Shot Medical Image Segmentation With Foundational Models
- Provable acceleration for diffusion models under minimal assumptions
- Prover-Verifier Games improve legibility of LLM outputs
- Pushing the Limits of Large Language Model Quantization via the Linearity Theorem
- Q*: Improving Multi-step **Reasoning** for LLMs with Deliberative Planning
- Quantitative Assessment of Intersectional Empathetic Bias and Understanding
- **Quantum** Attention for Vision Transformers in High Energy Physics
- Quark: Real-time, High-resolution, and General Neural View Synthesis
- Qwen2.5-Coder Technical Report
- Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution
- Rainbow Teaming: Open-Ended Generation of Diverse Adversarial Prompts
- RAVEN: In-Context Learning with Retrieval-Augmented Encoder-Decoder Language Models
- RE-Bench: Evaluating frontier AI R&D capabilities of language model **agents** against human experts
- Recent Advances in Attack and Defense Approaches of Large Language Models
- Recursive Introspection: Teaching Language Model Agents How to **Self-Improve**
- Recycled Attention: Efficient **inference** for long-context language models
- Reducing the Barriers to Entry for Foundation Model Training
- REDUCIO! Generating 1024×1024 Video within 16 Seconds using Extremely Compressed Motion Latents
- Refusal Tokens: A Simple Way to Calibrate Refusals in Large Language Models
- Reminding **Multimodal** Large Language Models of Object-aware Knowledge with Retrieved Tags
- RepairAgent: An Autonomous, LLM-Based **Agent** for Program Repair
- Representation Shattering in Transformers: A Synthetic Study with Knowledge Editing
- Reprogramming Foundational Large Language Models(LLMs) for Enterprise Adoption for Spatio-Temporal Forecasting Applications: Unveiling a New Era in Copilot-Guided **Cross-Modal** Time Series Representation Learning
- Rethinking Token Reduction in MLLMs: Towards a Unified Paradigm for Training-Free Acceleration

- Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach
- Retrieval with Learned Similarities
- RetrievalAttention: Accelerating Long-Context LLM **Inference** via Vector Retrieval
- Retrieving **Semantics** from the Deep: an RAG Solution for Gesture Synthesis
- ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation
- ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization
- RoboGolf: Mastering Real-World Minigolf with a Reflective **Multi-Modality** Vision-Language Model
- RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics
- Robot Navigation Using Physically Grounded Vision-Language Models in Outdoor Environments
- Robotic Control via **Embodied** Chain-of-Thought **Reasoning**
- Robots Can Multitask Too: Integrating a **Memory** Architecture and LLMs for Enhanced Cross-Task Robot Action Generation
- RuleAlign: Making Large Language Models Better Physicians with Diagnostic Rule Alignment
- Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?
- SAM 2: Segment Anything in Images and Videos
- SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation
- SAMIC: Segment Anything with In-Context Spatial Prompt Engineering
- SANER: Annotation-free Societal Attribute Neutralizer for Debiasing CLIP
- SARO: Space-Aware Robot System for Terrain Crossing via Vision-Language Model
- Scalable and Accurate Graph **Reasoning** with LLM-based **Multi-Agents**
- Scaling Cross-**Embodied** Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation
- Scaling Granite Code Models to 128K Context
- Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers
- Scaling Robot Policy Learning via Zero-Shot Labeling with Foundation Models
- Scaling Speech-Text Pre-training with Synthetic Interleaved Data
- Scaling up Masked Diffusion Models on Text
- Scaling Wearable Foundation Models
- SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding
- Score-based generative diffusion with "active" correlated noise sources
- SeedLM: Compressing LLM Weights into Seeds of Pseudo-Random Generators
- SegPoint: Segment Any Point Cloud via Large Language Model
- SELA: Tree-Search Enhanced LLM **Agents** for Automated Machine Learning
- Self-Taught Optimizer (STOP): Recursively **Self-Improving** Code Generation
- **Semantically**-Driven Disambiguation for Human-Robot Interaction
- SepLLM: Accelerate Large Language Models by Compressing One Segment into One Separator
- SGFormer: Single-Layer Graph Transformers with Approximation-Free Linear Complexity

- SIEVE: General Purpose Data Filtering System Matching GPT-4o Accuracy at 1% the Cost
- Simpler Diffusion (SiD2): 1.5 FID on ImageNet512 with pixel-space diffusion
- SimPO: Simple Preference Optimization with a Reference-Free Reward
- SimTube: Generating Simulated Video Comments through **Multimodal** AI and User Personas
- SkillMimicGen: Automated Demonstration Generation for Efficient Skill Learning and Deployment
- Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families
- Small Molecule Optimization with Large Language Models
- Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling
- SMILE: Zero-Shot Sparse **Mixture** of Low-Rank Experts Construction From Pre-Trained Foundation Models
- Software Performance Engineering for Foundation Model-Powered Software (FMware)
- SoK: Watermarking for AI-Generated Content
- Solving Robotics Problems in Zero-Shot with Vision-Language Models
- Source2Synth: Synthetic Data Generation and Curation Grounded in Real Data Sources
- SpaceMesh: A Continuous Representation for Learning Manifold Surface Meshes
- SparQ Attention: Bandwidth-Efficient LLM Inference
- sPhinX: Sample Efficient Multilingual Instruction Fine-Tuning Through N-shot Guided Prompting
- Spider: Any-to-Many **Multimodal** LLM
- Spider2-V: How Far Are **Multimodal Agents** From Automating Data Science and Engineering Workflows?
- SPRIG: Improving Large Language Model Performance by System Prompt Optimization
- Squeezed Attention: Accelerating Long Context Length LLM **Inference**
- Stable Audio Open
- Stateful Large Language Model Serving with Pensieve
- Steering Masked Discrete Diffusion Models via Discrete Denoising Posterior Prediction
- Stepping on the Edge: Curvature Aware Learning Rate Tuners
- Stronger Random Baselines for In-Context Learning
- StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions
- Surveying the space of descriptions of a composite system with machine learning
- SVDQuant: Absorbing Outliers by Low-Rank Components for 4-Bit Diffusion Models
- SWAN: Preprocessing SGD Enables Adam-Level Performance On LLM Training With Significant Memory Reduction
- Symmetry-Enriched Learning: A Category-Theoretic Framework for Robust Machine Learning Models
- SynCode: LLM Generation with Grammar Augmentation
- SynEHRgy: Synthesizing Mixed-Type Structured Electronic Health Records using Decoder-Only Transformers
- SynthBA: Reliable Brain Age Estimation Across Multiple MRI Sequences and Resolutions

- Synthetic continued pretraining
- SynthVLM: High-Efficiency and High-Quality Synthetic Data for Vision Language Models
- T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation
- TableGPT2: A Large **Multimodal** Model with Tabular Data Integration
- TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling
- Tailor3D: Customized 3D Assets Editing and Generation with Dual-Side Images
- Takin: A Cohort of Superior Quality Zero-shot Speech Generation Models
- TALK-Act: Enhance Textural-Awareness for 2D Speaking Avatar Reenactment with Diffusion Model
- Taming Data and Transformers for Audio Generation
- Targeting the Core: A Simple and Effective Method to Attack RAG-based **Agents** via Direct LLM Manipulation
- Teaching Transformers Causal **Reasoning** through Axiomatic Training
- Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs
- TextHawk2: A Large Vision-Language Model Excels in Bilingual OCR and Grounding with 16x Fewer Tokens
- The Art of Saying No: Contextual Noncompliance in Language Models
- The Base-Rate Effect on LLM Benchmark Performance: Disambiguating Test-Taking Strategies from Benchmark Performance
- The Future of Large Language Model Pre-training is Federated
- The Future of Software Testing: AI-Powered Test Case Generation and Validation
- The Matrix: Infinite-Horizon **World Generation** with Real-Time Moving Control
- The opportunities and risks of large language models in mental health
- The Prompt Report: A Systematic Survey of Prompting Techniques
- The **Semantic** Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities
- The Tug-of-War Between Deepfake Generation and Detection
- TheAgentCompany: Benchmarking LLM **Agents** on Consequential Real World Tasks
- Theia: Distilling Diverse Vision Foundation Models for Robot Learning
- Thinking in Space: How **Multimodal** Large Language Models See, Remember, and Recall Spaces
- To CoT or not to CoT? Chain-of-thought helps mainly on math and **symbolic reasoning**
- Token Statistics Transformer: Linear-Time Attention via Variational Rate Reduction
- TOMATO: Assessing Visual Temporal **Reasoning** Capabilities in **Multimodal** Foundation Models
- Topology-Based Reconstruction Prevention for Decentralised Learning
- Topology-guided Hypergraph Transformer Network: Unveiling Structural Insights for Improved Representation
- Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements

- Towards CausalGPT: A **Multi-Agent** Approach for Faithful **Knowledge Reasoning** via Promoting Causal Consistency in LLMs
- Towards Effective and Efficient Continual Pre-training of Large Language Models
- Towards End-to-End Open Conversational Machine Reading
- Towards Foundation Models for 3D Vision: How Close Are We?
- Towards Foundation-model-based **Multiagent** System to Accelerate AI for Social Impact
- Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models
- Towards Leveraging Contrastively Pretrained Neural Audio Embeddings for Recommender Tasks
- Towards Low-bit Communication for Tensor Parallel LLM **Inference**
- Towards Maximum Likelihood Training for Transducer-based Streaming Speech Recognition
- Towards Unifying Understanding and Generation in the Era of Vision Foundation Models: A Survey from the Autoregression Perspective
- Towards a Classification of Open-Source ML Models and Datasets for Software Engineering
- Training Large Language Models to **Reason** in a Continuous Latent Space
- Training-free Regional Prompting for Diffusion Transformers
- TRANSAGENT: An LLM-Based **Multi-Agent** System for Code Translation
- Transferable Ensemble Black-box Jailbreak Attacks on Large Language Models
- Transformers are Deep Optimizers: Provable In-Context Learning for Deep Model Training
- Transformers Can Do Bayesian **Inference**
- Transformers to SSMs: Distilling Quadratic Knowledge to Subquadratic Models
- Transformers Use Causal **World Models** in Maze-Solving Tasks
- Transforming the Hybrid Cloud for Emerging AI Workloads
- Transfusion: Predict the Next Token and Diffuse Images with One **Multi-Modal** Model
- TÜLU 3: Pushing Frontiers in Open Language Model Post-Training
- Tuning-free coreset Markov chain Monte Carlo
- TurboAttention: Efficient Attention Approximation For High Throughputs LLMs
- Turn Every Application into an **Agent**: Towards Efficient Human-Agent-Computer Interaction with API-First LLM-Based Agents
- TurtleBench: Evaluating Top Language Models via Real-World Yes/No Puzzles
- UADA3D: Unsupervised Adversarial Domain Adaptation for 3D Object Detection with Sparse LiDAR and Large Domain Gaps
- UKAN: Unbound Kolmogorov-Arnold Network Accompanied with Accelerated Library
- UltraEval: A Lightweight Platform for Flexible and Comprehensive Evaluation for LLMs
- Uncovering LLM-Generated Code: A Zero-Shot Synthetic Code Detector via Code Rewriting
- Universal Sound Separation with Self-Supervised Audio Masked Autoencoder
- Unlocking Guidance for Discrete State-Space Diffusion and Flow Models
- Unlocking the Power of Gradient Guidance for Structure-Based Molecule Optimization

- Unraveling **Cross-Modality** Knowledge Conflict in Large Vision-Language Models
- UrbanWorld: An Urban **World Model** for 3D City Generation
- Using LLM for Real-Time Transcription and Summarization of Doctor-Patient Interactions into ePuskesmas in Indonesia
- Utilizing Large Language Models to Synthesize Product Desirability Datasets
- VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models
- VFA: Vision Frequency Analysis of Foundation Models and Human
- Video-STaR: Self-Training Enables Video Instruction Tuning with Any Supervision
- VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation
- Virchow2: Scaling Self-Supervised **Mixed** Magnification **Models** in Pathology
- VIRUS-NeRF -- Vision, InfraRed and UltraSonic based Neural Radiance Fields
- VisualPredicator: Learning Abstract **World Models** with **Neuro-Symbolic** Predicates for Robot Planning
- VLMEvalKit: An Open-Source Toolkit for Evaluating Large **Multi-Modality** Models
- VPTQ: Extreme Low-bit Vector Post-Training Quantization for Large Language Models
- VRSD: Rethinking Similarity and Diversity for Retrieval in Large Language Models
- Warped Diffusion: Solving Video Inverse Problems with Image Diffusion Models
- WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling
- When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models
- When AI Meets Finance (StockAgent): Large Language Model-based Stock Trading in Simulated Real-world Environments
- When Backdoors Speak: Understanding LLM Backdoor Attacks Through Model-Generated Explanations
- Will Large Language Models be a Panacea to Autonomous Driving?
- Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy
- Wolf: Captioning Everything with a World Summarization Framework
- Words2Contact: Identifying Support Contacts from Verbal Instructions Using Foundation Models
- WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines
- xGen-MM (BLIP-3): A Family of Open Large **Multimodal** Models
- xGen-VideoSyn-1: High-fidelity Text-to-Video Synthesis with Compressed Representations
- XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models
- You Name It, I Run It: An LLM **Agent** to Execute Tests of Arbitrary Projects
- Your **Mixture**-of-Experts LLM Is Secretly an Embedding Model For Free
- ZS4C: Zero-Shot Synthesis of Compilable Code for Incomplete Code Snippets using LLMs
- π0: A Vision-Language-Action Flow Model for General Robot Control

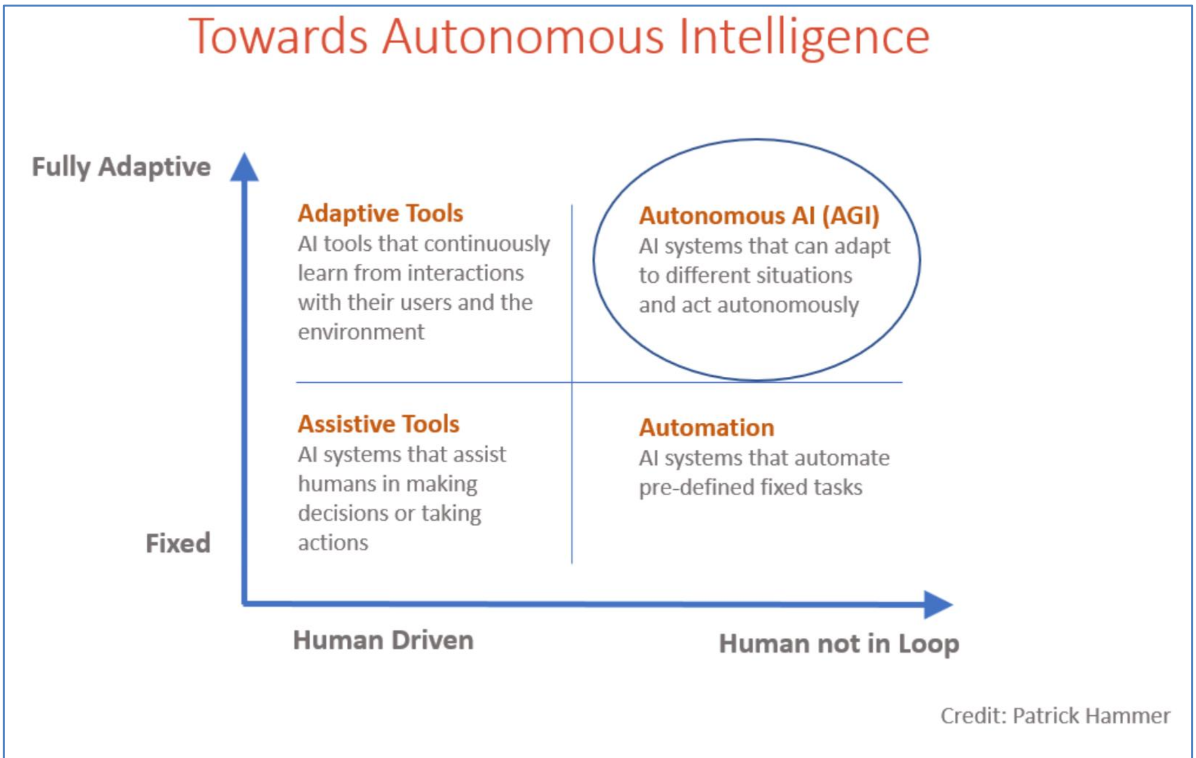
Key issues relevant to our Project

- **Agency, Embodiment, Edge (terminal) devices etc.**
- **Multi-agent systems MAS, Mixture of Experts MoE, cooperation, collaboration etc.**
- **Multi-modality & Cross-modality etc.**
- **Inference & Reasoning etc.**
- **Adaptation, Evolution, Self-improvement, SO etc.**
- **Memory, Knowledge Graphs KG etc.**
- **Quantum, Quntization etc.**
- **Semantic, Symbolic etc.**
- **World Models**

Q. New Findings in 2024 H2+

Addition info (here illustrations only) for some papers from Chapter 62

- [Voss & Jovanovic (2023b)] - **Why We Don't Have AGI Yet**



Towards Autonomous Intelligence

Fully Adaptive

Fixed

Human Driven

Human not in Loop

Adaptive Tools
AI tools that continuously learn from interactions with their users and the environment

Autonomous AI (AGI)
AI systems that can adapt to different situations and act autonomously

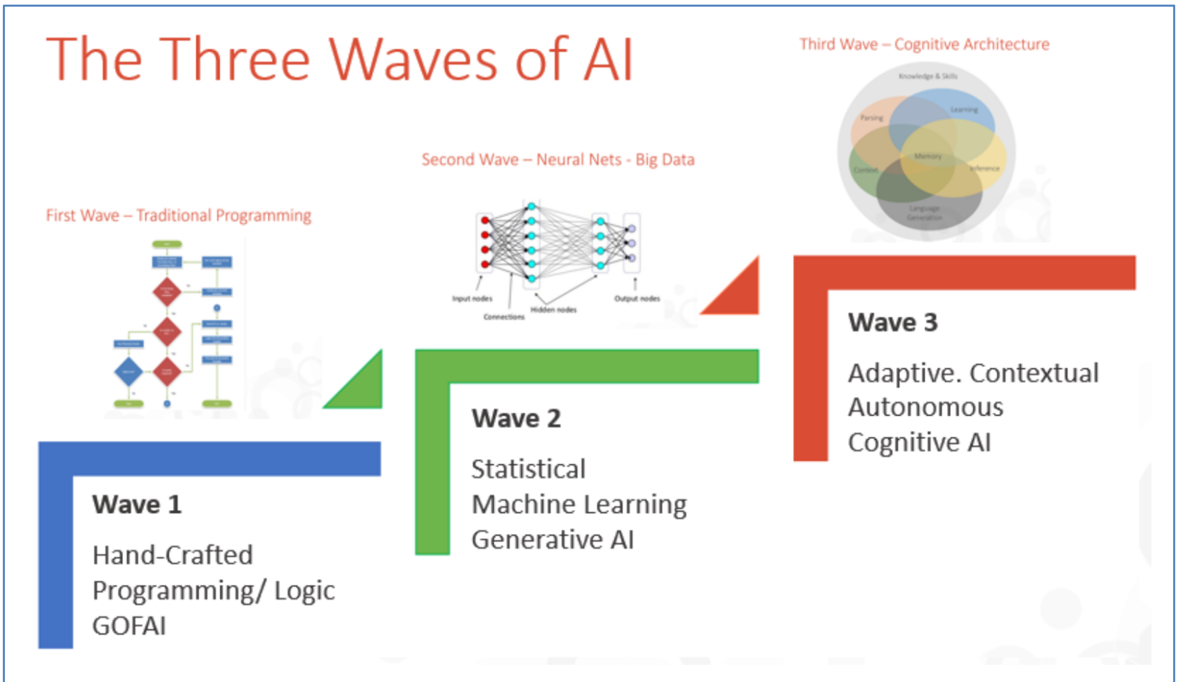
Assistive Tools
AI systems that assist humans in making decisions or taking actions

Automation
AI systems that automate pre-defined fixed tasks

Credit: Patrick Hammer

Figure 2. Dimensions of Adaptive Autonomous Intelligence.

{Terms for Index - Adaptive Autonomous AI)



The Three Waves of AI

First Wave – Traditional Programming

Second Wave – Neural Nets - Big Data

Third Wave – Cognitive Architecture

Wave 1
Hand-Crafted Programming/ Logic GOFAI

Wave 2
Statistical Machine Learning Generative AI

Wave 3
Adaptive. Contextual Autonomous Cognitive AI

Figure 3. The Three waves of AI. Adapted from DARPA [11].

- [Sukhobokov et al. (2024)] **- A Universal Knowledge Model and Cognitive Architecture for Prototyping AGI** *{Metagraph, Memory, Consciousness)*

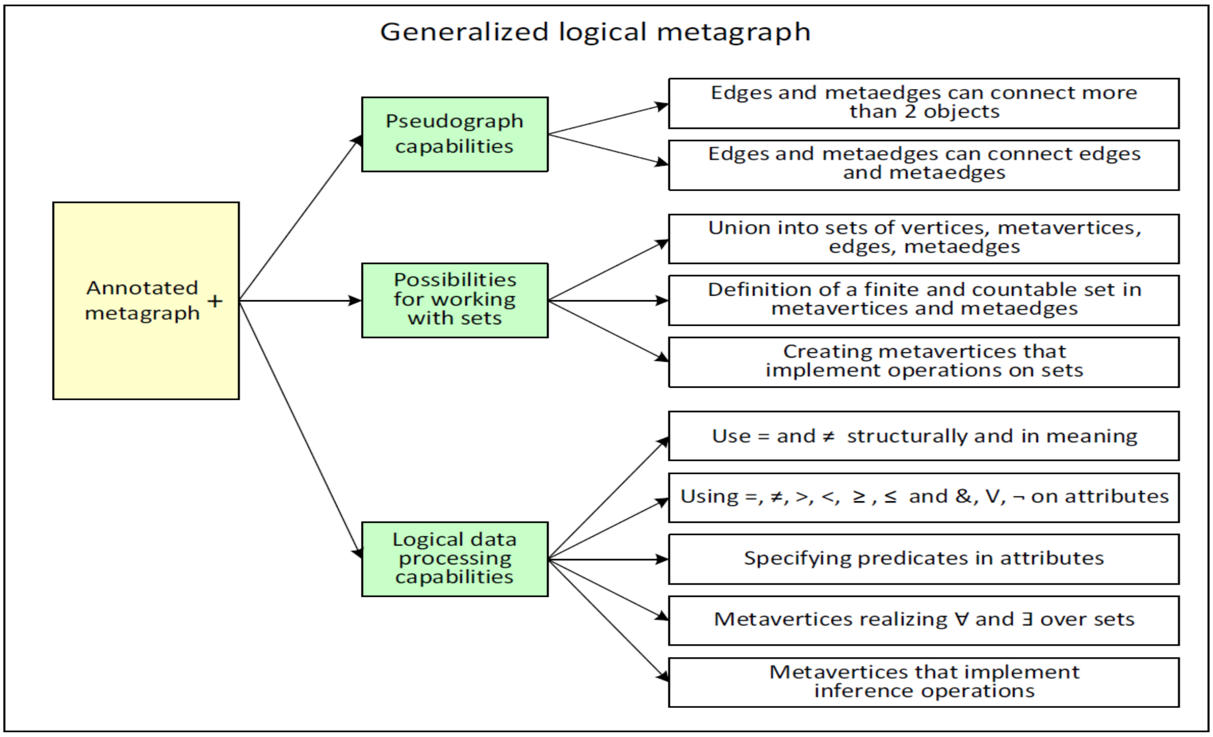

Generalized logical metagraph
Annotated metagraph +
Pseudograph capabilities
Possibilities for working with sets
Logical data processing capabilities
Edges and metaedges can connect more than 2 objects
Edges and metaedges can connect edges and metaedges
Union into sets of vertices, metavertices, edges, metaedges
Definition of a finite and countable set in metavertices and metaedges
Creating metavertices that implement operations on sets
Use = and ≠ structurally and in meaning
Using =, ≠, >, <, ≥ , ≤ and &, V, ¬ on attributes
Specifying predicates in attributes
Metavertices realizing ∀ and ∃ over sets
Metavertices that implement inference operations

Fig. 1. An expansion of annotated metagraph.

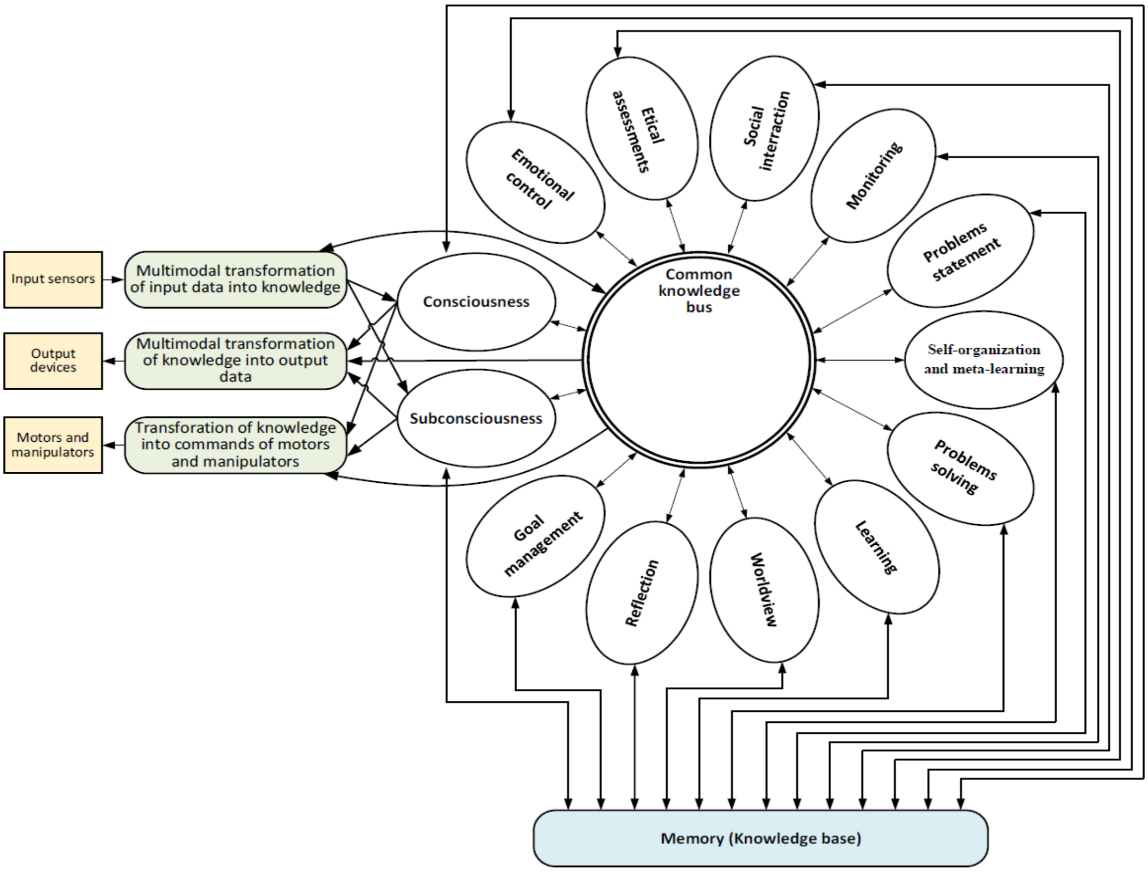
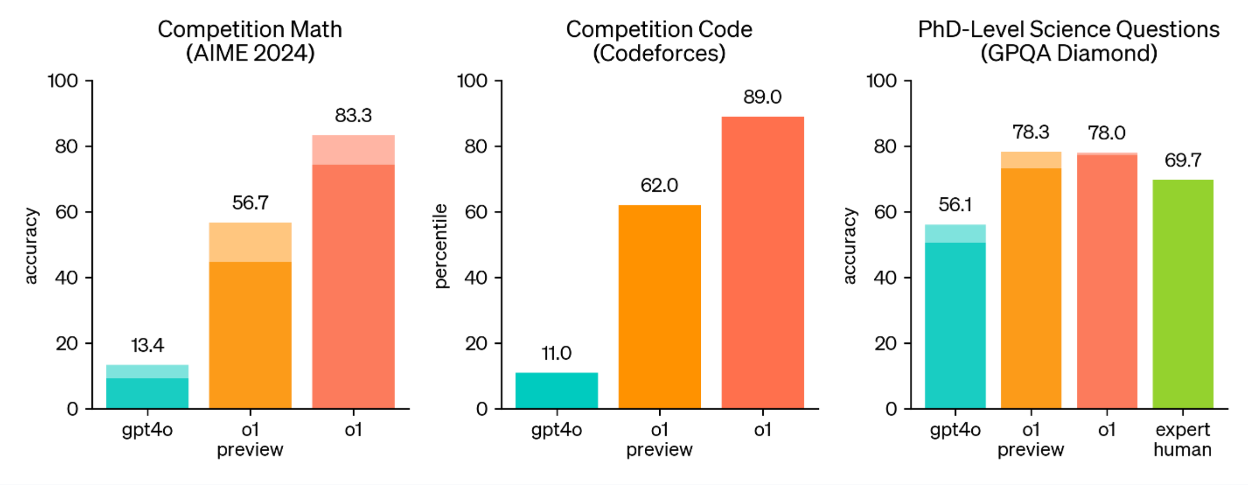
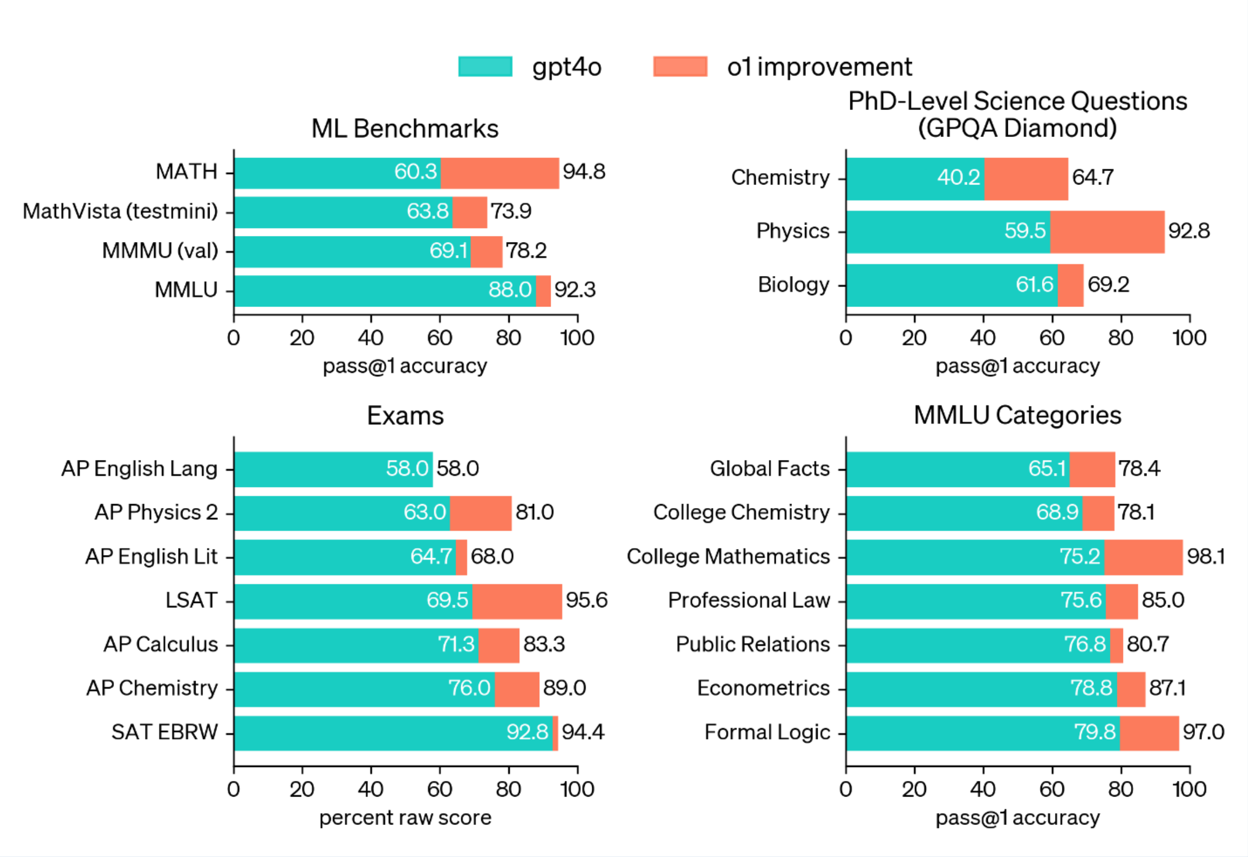


Fig. 2. A diagram of the cognitive architecture that can be used to develop AGI prototypes.

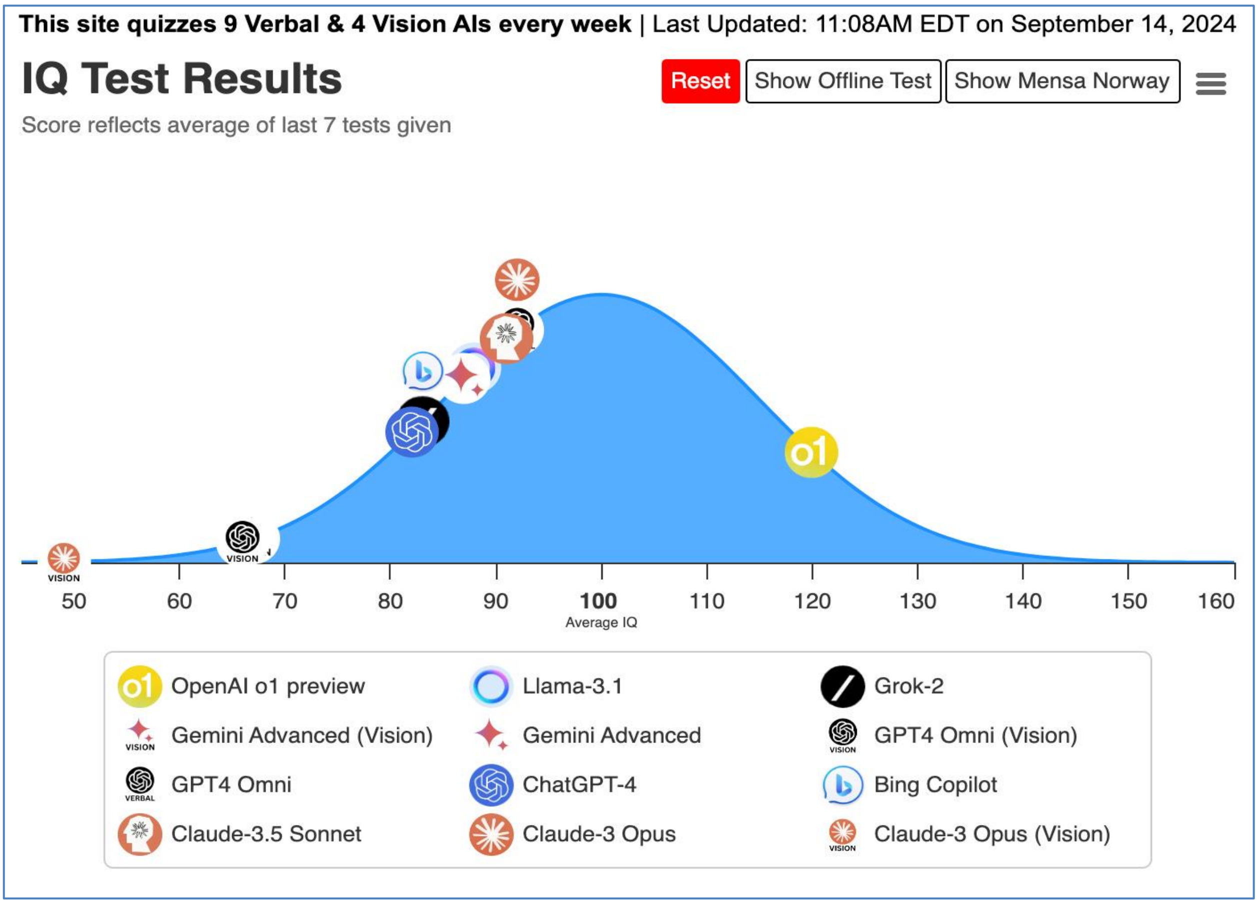
- [OpenAI (2024)] - **Learning to Reason with LLMs**





{Benchmarks, Science, Mathematics, GPT}

- [Lott (2024)] **– LLMs IQ tests comparison** *{GPT, Gemini, Grok, Llama, Claude-3)*

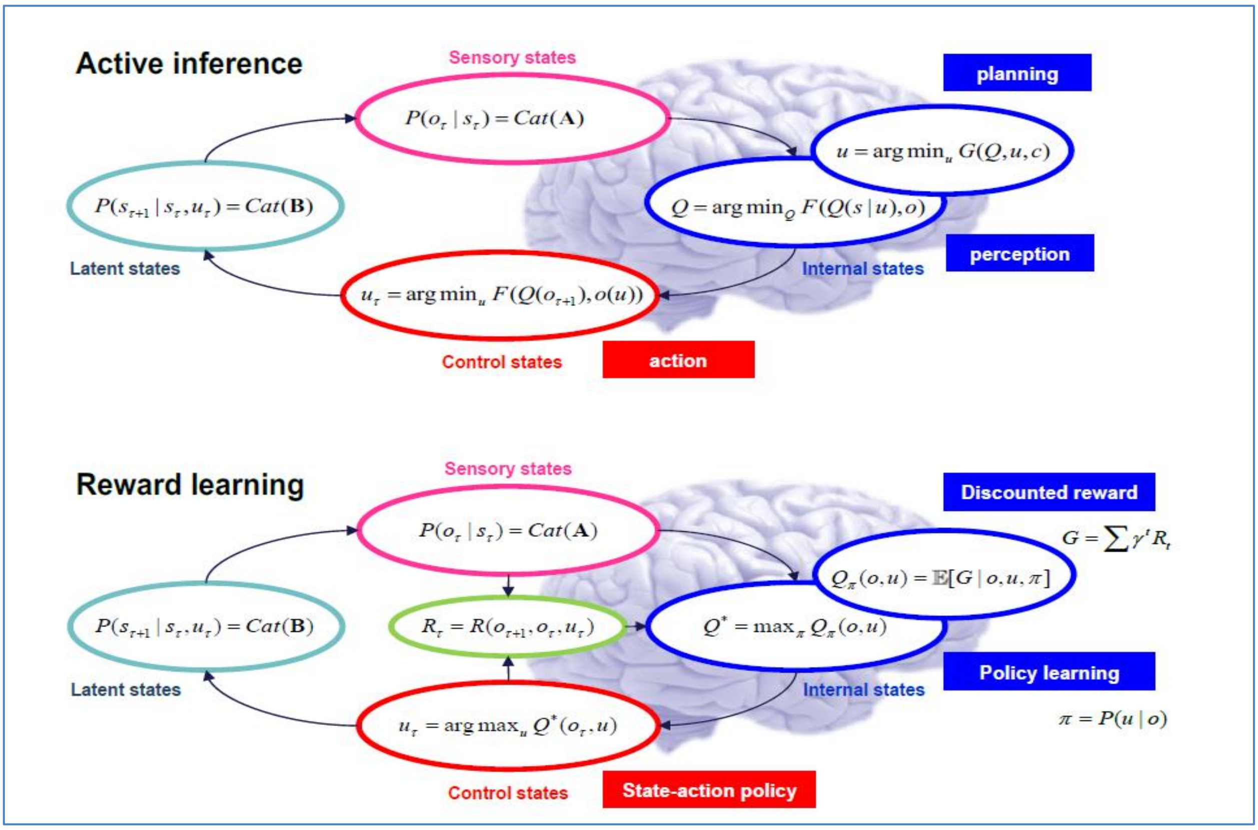


This site quizzes 9 Verbal & 4 Vision AIs every week | Last Updated: 11:08AM EDT on September 14, 2024

IQ Test Results

Score reflects average of last 7 tests given

- [Friston et al. (2024)] - **From pixels to planning: scale-free active inference**



Active inference
Sensory states
$P(o_\tau \mid s_\tau) = Cat(\mathbf{A})$
planning
$u = \arg\min_u G(Q,u,c)$
$P(s_{\tau+1} \mid s_\tau, u_\tau) = Cat(\mathbf{B})$
$Q = \arg\min_Q F(Q(s \mid u), o)$
perception
Latent states
Internal states
$u_\tau = \arg\min_u F(Q(o_{\tau+1}), o(u))$
Control states
action
Reward learning
Sensory states
Discounted reward
$P(o_\tau \mid s_\tau) = Cat(\mathbf{A})$
$G = \sum \gamma^t R_t$
$Q_\pi(o,u) = \mathbb{E}[G \mid o,u,\pi]$
$P(s_{\tau+1} \mid s_\tau, u_\tau) = Cat(\mathbf{B})$
$R_\tau = R(o_{\tau+1}, o_\tau, u_\tau)$
$Q^* = \max_\pi Q_\pi(o,u)$
Policy learning
Latent states
Internal states
$\pi = P(u \mid o)$
$u_\tau = \arg\max_u Q^*(o_\tau, u)$
Control states
State-action policy

- [Nelson (2024a)] - **Prioritization, Iteration, and Convergence in Cognitive Systems: Bayesian Inference, Perceptual Gating, and the Requirement Equation**

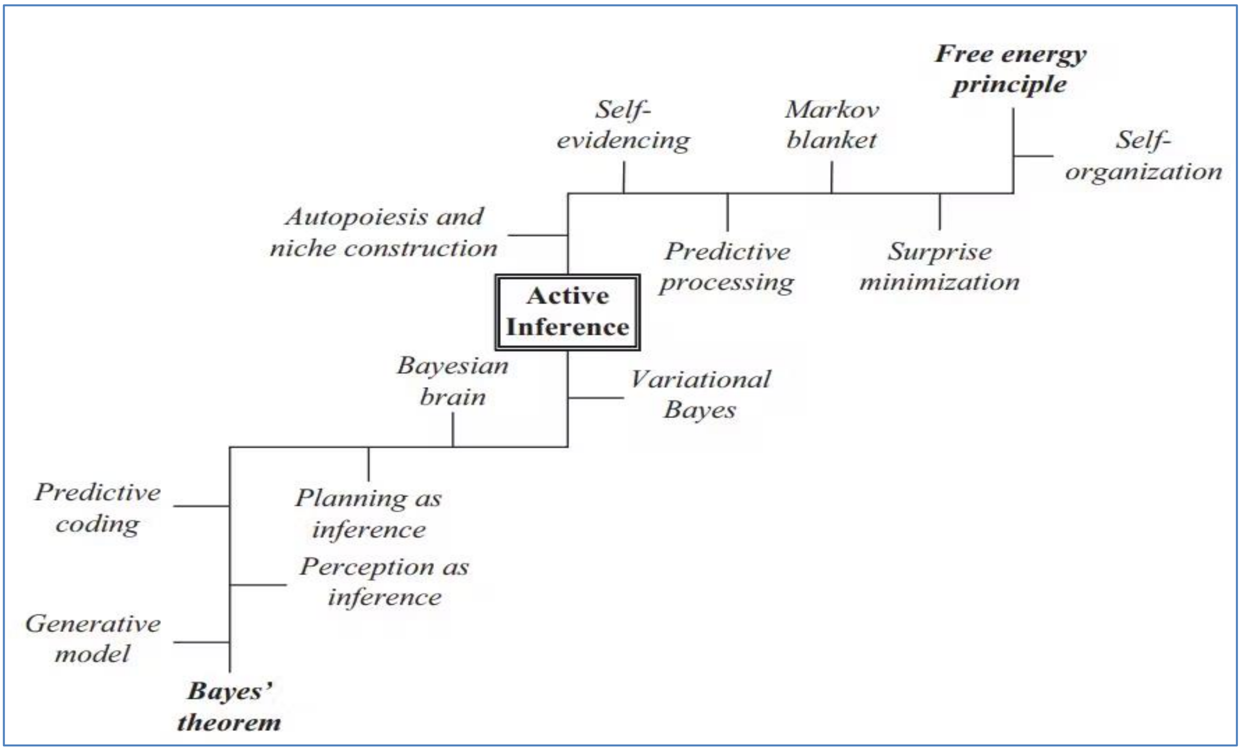


Figure 1: The High and Low Roads to Active Inference (from Parr et al. 2022)

{Free Energy, SO)

- [Da Costa et al. (2024a)] - **Possible principles for aligned structure learning agents**

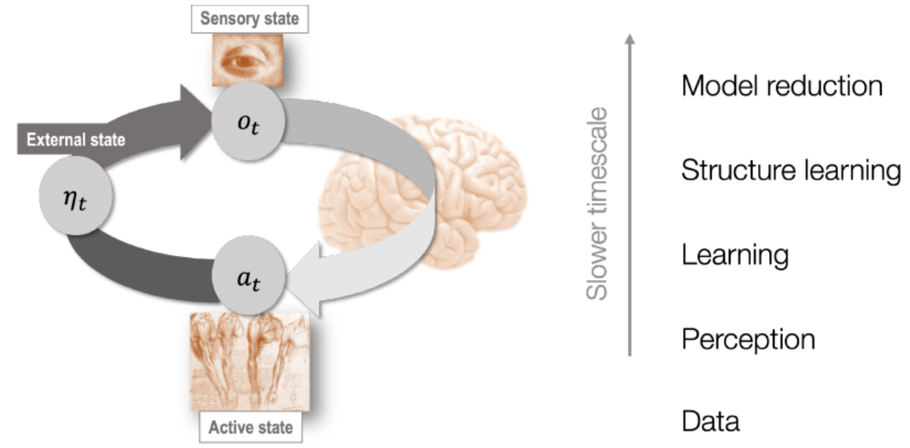


Figure 5: **Structure learning agents.** This figure summarizes the various processes underlying agents that learn causal structure. *Left:* The external state of the environment causes a sensory state, which is processed by the agent resulting in a choice of active state, which influences the next external state—and the perception action cycle repeats. *Right:* Agents have access to an incoming stream of data, and use this for perception, learning, structure learning and finally model reduction, which have a technical meaning here: namely, inferring the states, parameters, and structure of models, respectively. These processes unfold at slower and slower timescales: this is necessary since accurate inference about e.g. causal structure, requires many more data points than inference about parameters or states, under a causal structure.

{Model, Data)

- [Zhang et al. (2024)] - **Intelligence at the Edge of Chaos** *{LLM, pre-training)*

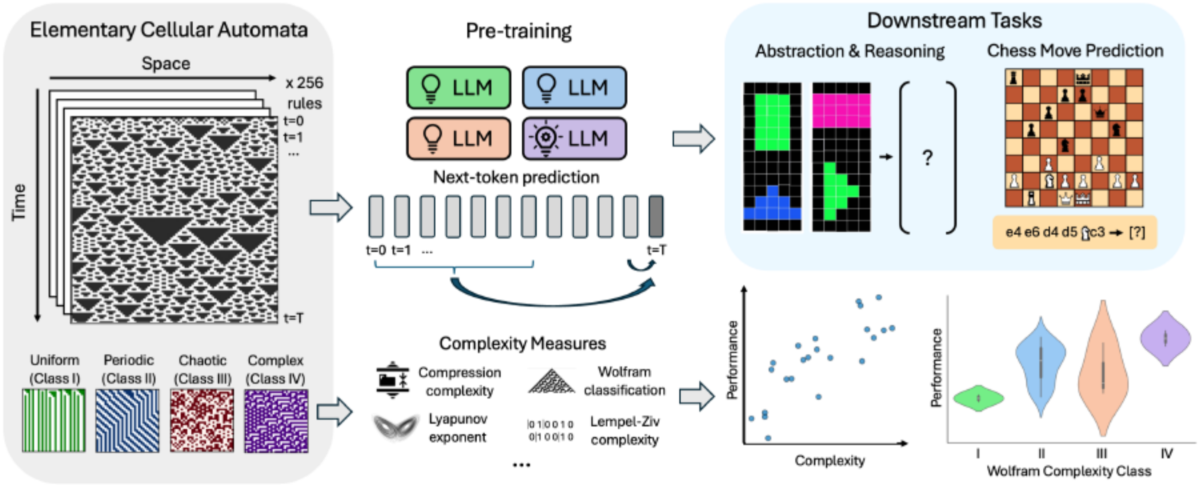


Figure 1: Our framework for investigating the link between complexity and intelligence. We pretrain Large Language Models (LLMs) on Elementary Cellular Automata (ECAs) from different complexity classes using next-token prediction, then evaluate them on downstream reasoning and chess move prediction tasks. We use various measures to analyze the complexity of ECA rules, and quantify the relationship between complexity and downstream performance.

- [Hochberg (2024)] - **A Theory of Intelligences** *{Intelligence, Ecosystem)*

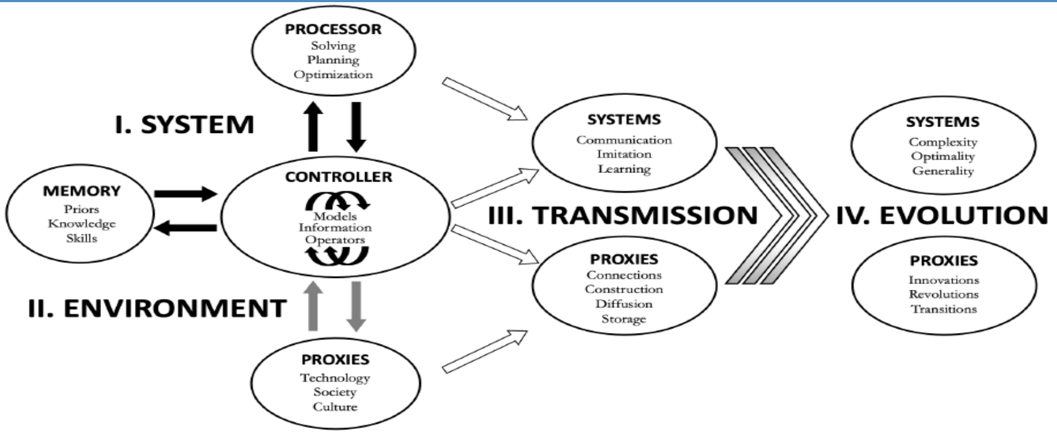


Figure 6. Schematic diagram of TIS in the intelligence ecosystem. The SYSTEM phenotype (I) is composed of the CONTROLLER, PROCESSOR and MEMORY. The CONTROLLER generates the goal-directed information based on generic abilities (Generic Operators, Table 1), and how these abilities are instrumentalized via the PROCESSOR, which addresses challenges in their difficulty and novelty, and the arbitration of exploration and exploitation, and in so doing, solves uncertainty and plans and optimizes paths to goal resolution. The SYSTEM interacts with (II) the ENVIRONMENT both in setting and addressing GOALS. The extended phenotypes are PROXIES such as social interactions, technology and culture. Intelligence can be codified *in* the SYSTEM as phenotypic traits, stored MEMORY, and hard-wired or plastic behaviors. Intelligence can also be codified *outside* the SYSTEM in (non-mutually exclusive) PROXIES, such as collectives, society, culture, artefacts, technology and institutions. The current SYSTEM (I and II) is based on past and current TRANSMISSION and intelligence trait EVOLUTION (not shown), and develops and integrates intelligence traits over the SYSTEM's lifetime (not shown), and influences future (III) TRANSMISSION and (IV) EVOLUTION. The intelligence niche is affected by SYSTEM EVOLUTION and possibly SYSTEM-dependent PROXY EVOLUTION and this occasionally produces intelligence innovations and, more rarely, intelligence transitions.

- [Cross et al. (2024)] - **Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models** *{MAS, LLM}*

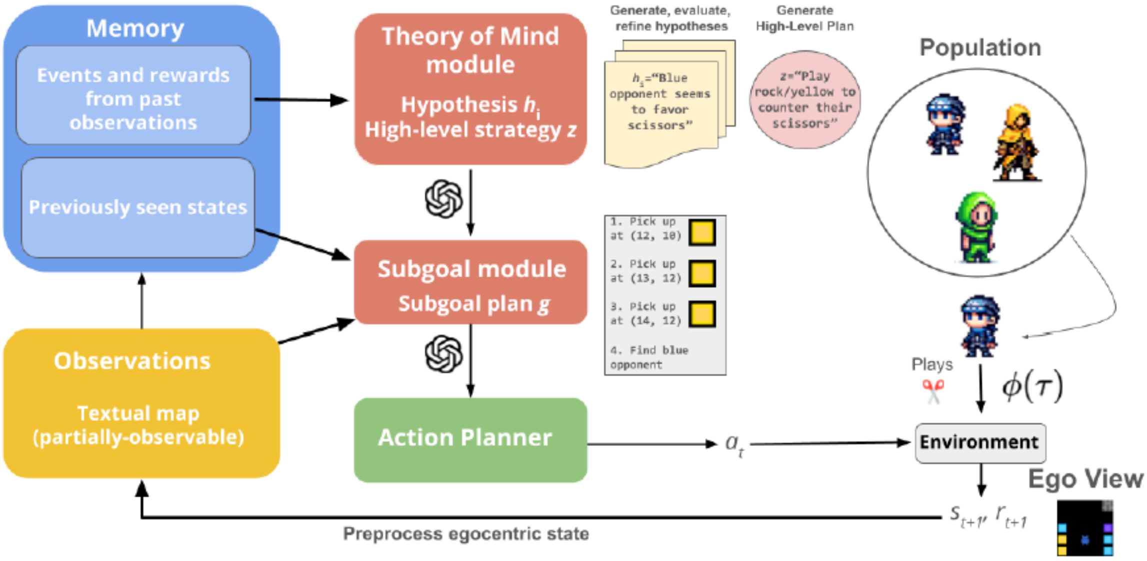


Figure 1: Hypothetical Minds architecture and depiction of model workflow

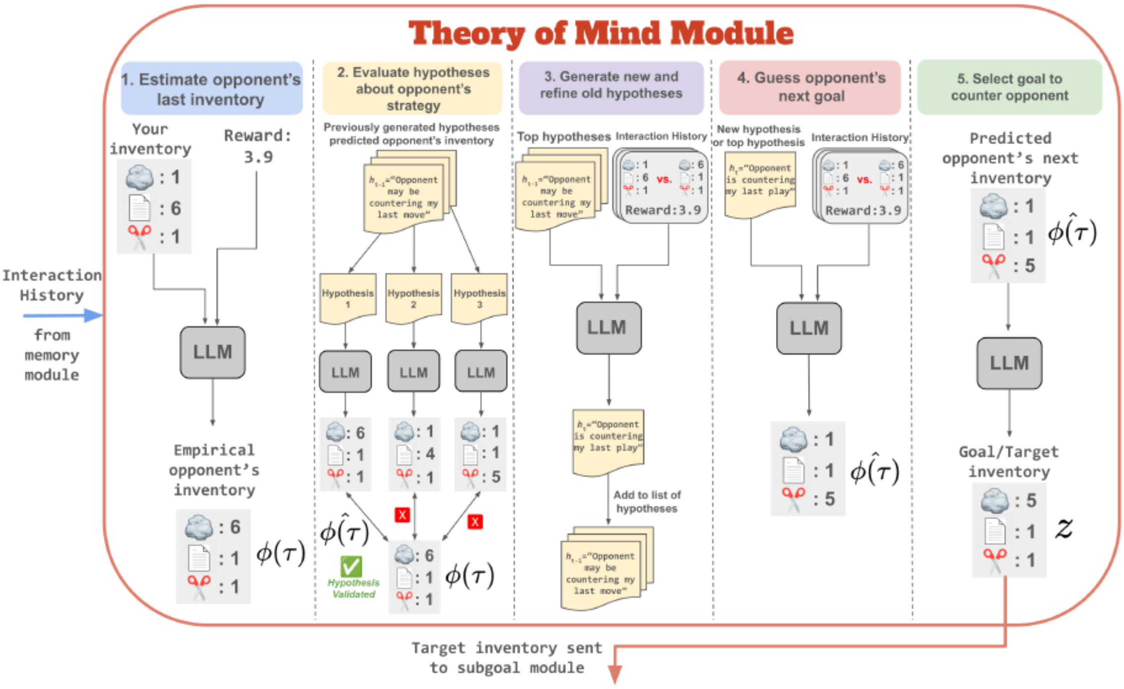

Theory of Mind Module
1. Estimate opponent's last inventory
2. Evaluate hypotheses about opponent's strategy
3. Generate new and refine old hypotheses
4. Guess opponent's next goal
5. Select goal to counter opponent
Interaction History
from memory module
Your inventory
: 1
: 6
: 1
Reward: 3.9
LLM
Empirical opponent's inventory
: 6
: 1
: 1
$\phi(\tau)$
Previously generated hypotheses predicted opponent's inventory
h_{t-1}="Opponent may be countering my last move"
Hypothesis 1
Hypothesis 2
Hypothesis 3
LLM
LLM
LLM
: 6
: 1
: 1
: 1
: 4
: 1
: 1
: 1
: 5
$\phi(\hat{\tau})$
Hypothesis Validated
: 6
: 1
: 1
$\phi(\tau)$
Top hypotheses
Interaction History
h_{t-1}="Opponent may be countering my last move"
: 1
: 6
: 1
vs.
: 6
: 1
: 1
Reward:3.9
LLM
h_t="Opponent is countering my last play"
Add to list of hypotheses
h_{t-1}="Opponent may be countering my last move"
New hypothesis or top hypothesis
Interaction History
h_t="Opponent is countering my last play"
: 1
: 6
: 1
vs.
: 6
: 1
: 1
Reward:3.9
LLM
: 1
: 1
: 5
$\phi(\hat{\tau})$
Predicted opponent's next inventory
: 1
: 1
: 5
$\phi(\hat{\tau})$
LLM
Goal/Target inventory
: 5
: 1
: 1
z
Target inventory sent to subgoal module

Figure 2: Theory of Mind (ToM) Module for Running With Scissors. This cognitive module receives input in the form of interaction history and outputs a target inventory as a goal for the subgoal module. Information is processed in 5 steps, including using the available information to generate, evaluate, and refine hypotheses about the opponent's strategy.

- [Danilenka et al. (2024)] - **Adaptive Active Inference Agents for Heterogeneous and Lifelong Federated Learning**

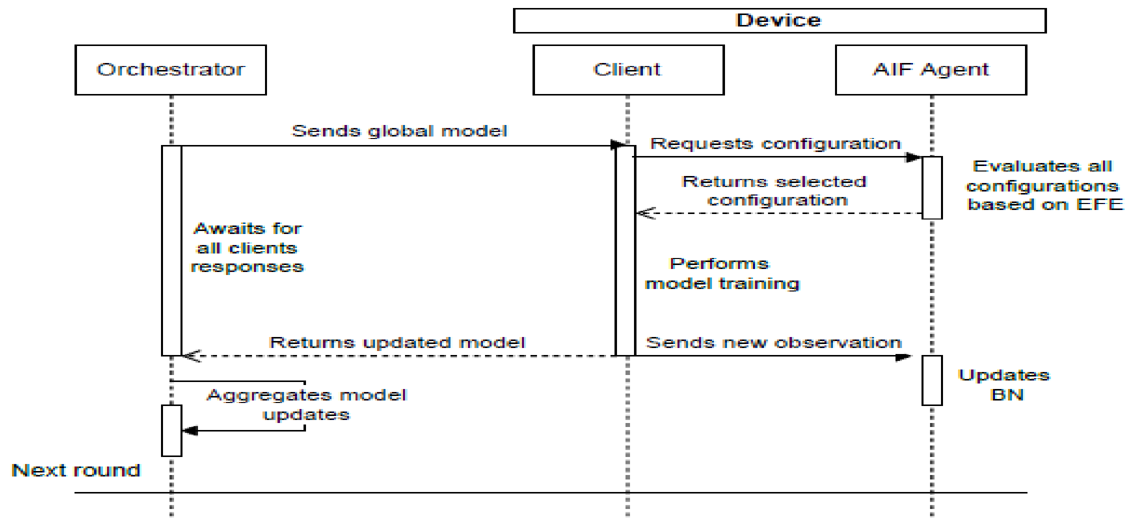


Fig. 2. Sequence diagram for one FL round of the proposed method

- [Arslan (2024)] - **Artificial Human Intelligence: The role of Humans in the Development of Next Generation AI**

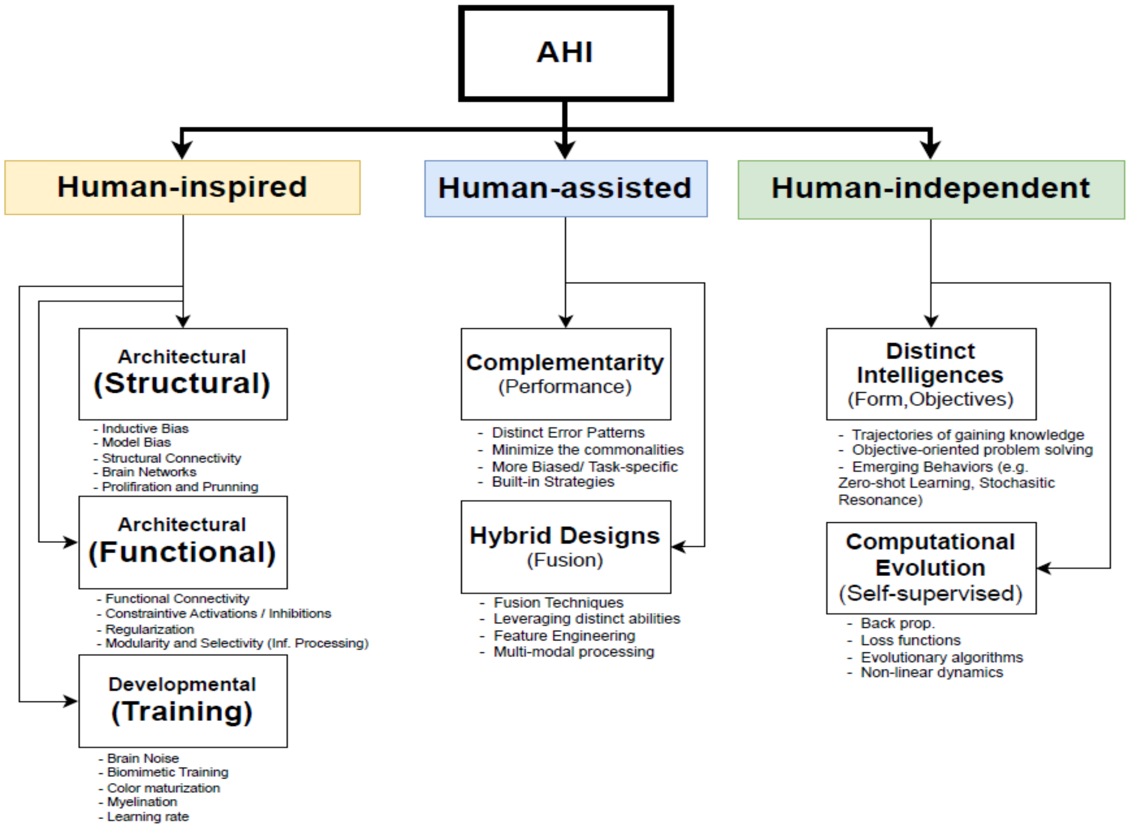

AHI
Human-inspired
Human-assisted
Human-independent
Architectural (Structural)
- Inductive Bias
- Model Bias
- Structural Connectivity
- Brain Networks
- Prolifiration and Prunning
Architectural (Functional)
- Functional Connectivity
- Constraintive Activations / Inhibitions
- Regularization
- Modularity and Selectivity (Inf. Processing)
Developmental (Training)
- Brain Noise
- Biomimetic Training
- Color maturization
- Myelination
- Learning rate
Complementarity (Performance)
- Distinct Error Patterns
- Minimize the commonalities
- More Biased/ Task-specific
- Built-in Strategies
Hybrid Designs (Fusion)
- Fusion Techniques
- Leveraging distinct abilities
- Feature Engineering
- Multi-modal processing
Distinct Intelligences (Form,Objectives)
- Trajectories of gaining knowledge
- Objective-oriented problem solving
- Emerging Behaviors (e.g. Zero-shot Learning, Stochasitic Resonance)
Computational Evolution (Self-supervised)
- Back prop.
- Loss functions
- Evolutionary algorithms
- Non-linear dynamics

Figure 3: Categorization of Human-centered intelligences and associated subcatagories.

- [Henriques et al. (2019)] - **The Tree of Knowledge System: A New Map for Big History**

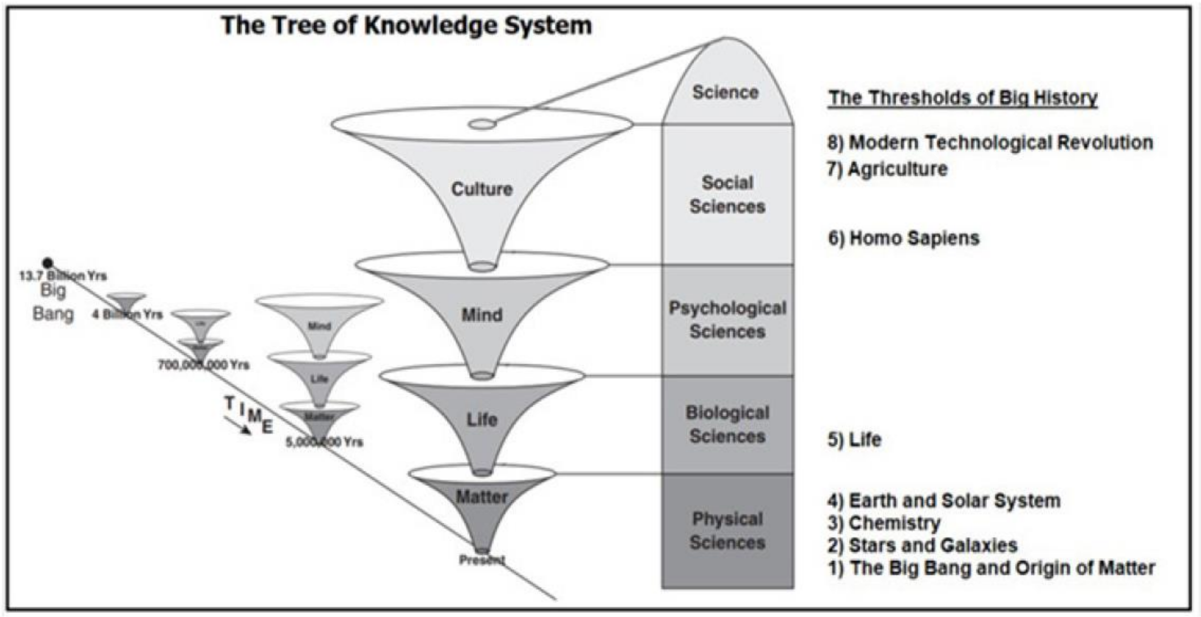


Figure 1. The Tree of Knowledge System and the Thresholds of Big History

{Matter, Life, Mind, Culture)

- [Chae et al. (2024)] - **Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation** *{Training)*

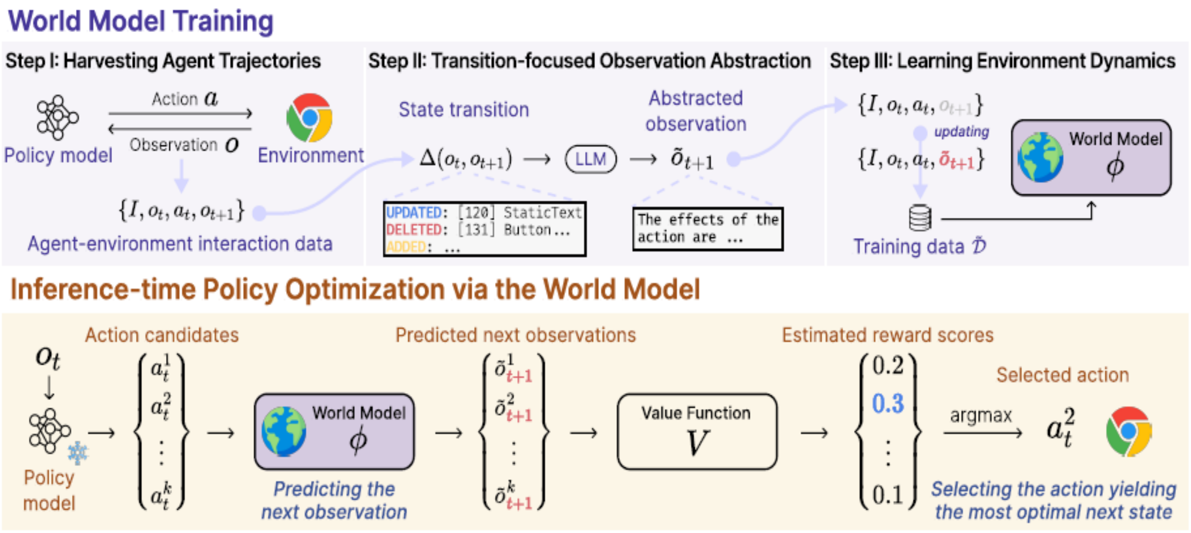

World Model Training
Step I: Harvesting Agent Trajectories
Action a
Observation o
Policy model
Environment
$\{I, o_t, a_t, o_{t+1}\}$
Agent-environment interaction data
Step II: Transition-focused Observation Abstraction
State transition
Abstracted observation
$\Delta(o_t, o_{t+1}) \longrightarrow$ LLM $\longrightarrow \tilde{o}_{t+1}$
UPDATED: [120] StaticText
DELETED: [131] Button...
ADDED: ...
The effects of the action are ...
Step III: Learning Environment Dynamics
$\{I, o_t, a_t, o_{t+1}\}$
updating
$\{I, o_t, a_t, \tilde{o}_{t+1}\}$
World Model ϕ
Training data $\mathcal{D}$
Inference-time Policy Optimization via the World Model
Action candidates
o_t
Policy model
a_t^1 a_t^2 $\vdots$ a_t^k
World Model ϕ
Predicting the next observation
Predicted next observations
$\tilde{o}_{t+1}^1$ $\tilde{o}_{t+1}^2$ $\vdots$ $\tilde{o}_{t+1}^k$
Value Function V
Estimated reward scores
0.2
0.3
$\vdots$
0.1
Selected action
argmax
a_t^2
Selecting the action yielding the most optimal next state

Figure 3: Framework overview. We first collect training data for world models (top). After training, we perform policy optimization by selecting the action leading to an optimal next state (bottom).

- [Christakopoulou, Mourad & Matari´c (2024)] - **Agents Thinking Fast and Slow: A Talker-Reasoner Architecture** *{MAS)*

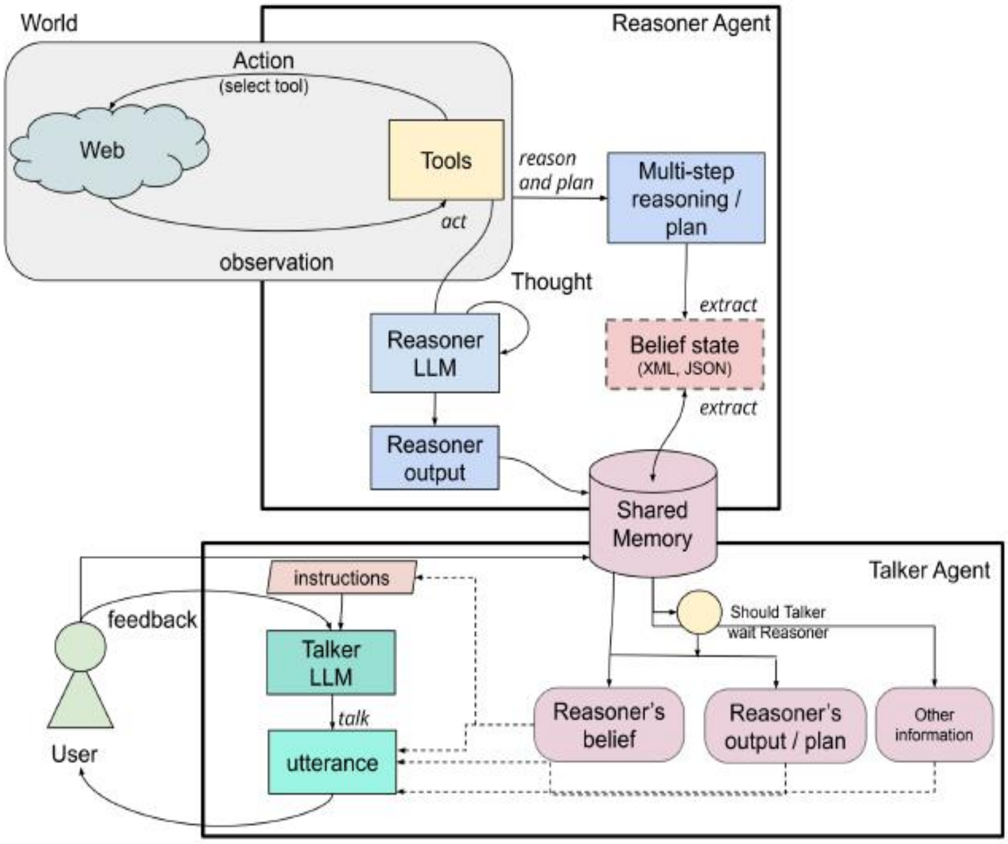


Figure 3: Diagram of Talker-Reasoner architecture.

- [Zhao Feifei et al. (2024)] - **Building Altruistic and Moral AI Agent with Brain-inspired Affective Empathy Mechanisms**

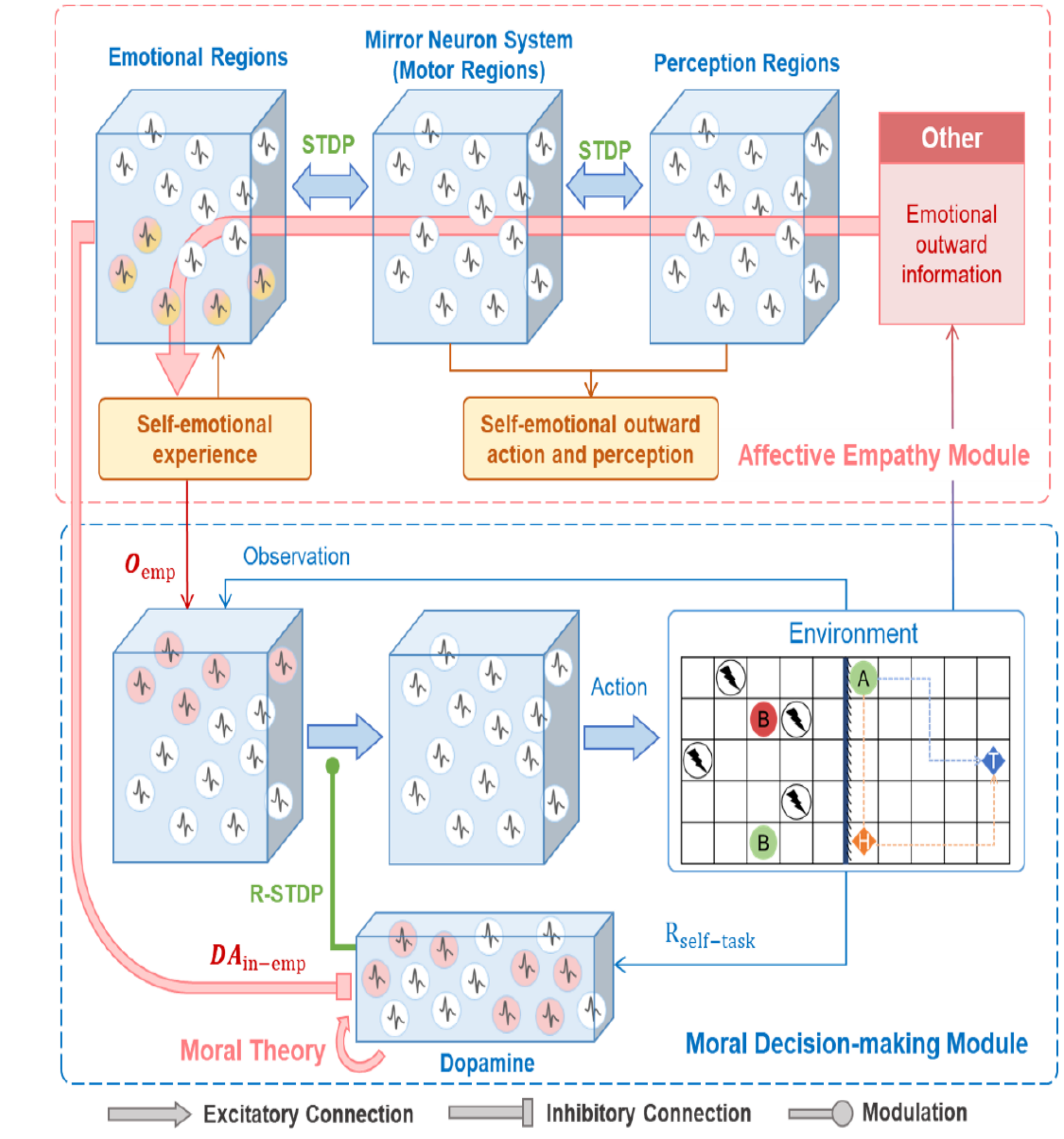


Fig. 1. The procedure of brain-inspired affective empathy-driven moral decision-making algorithm.

- [Microsoft (2024)] - **Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks** *{MAS)*

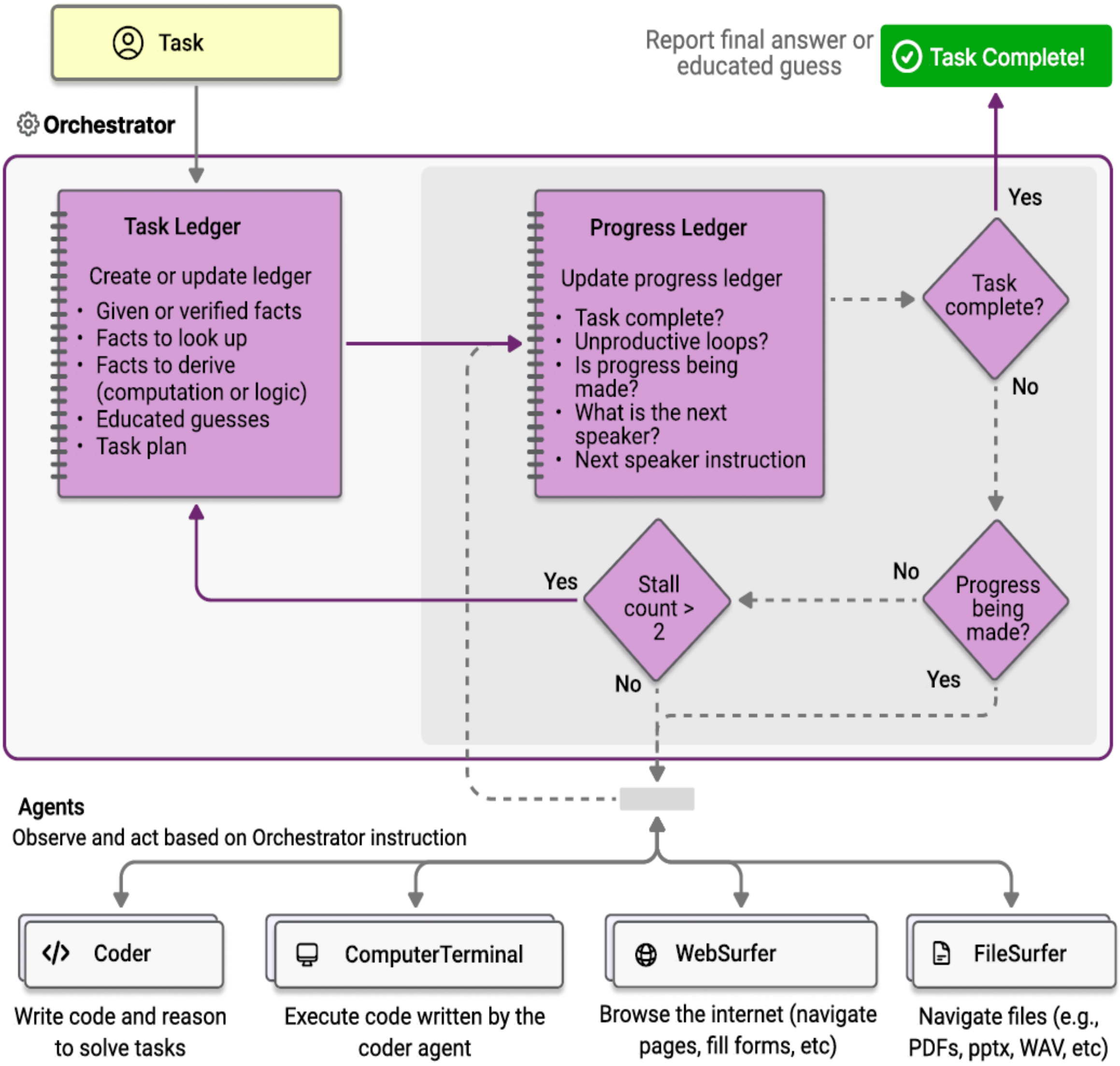

Task
Orchestrator
Report final answer or educated guess
Task Complete!
Task Ledger
Create or update ledger
• Given or verified facts
• Facts to look up
• Facts to derive (computation or logic)
• Educated guesses
• Task plan
Progress Ledger
Update progress ledger
• Task complete?
• Unproductive loops?
• Is progress being made?
• What is the next speaker?
• Next speaker instruction
Yes
Task complete?
No
Yes
Stall count > 2
No
No
Progress being made?
Yes
Agents
Observe and act based on Orchestrator instruction
Coder
Write code and reason to solve tasks
ComputerTerminal
Execute code written by the coder agent
WebSurfer
Browse the internet (navigate pages, fill forms, etc)
FileSurfer
Navigate files (e.g., PDFs, pptx, WAV, etc)

Figure 2: Magentic-One features an Orchestrator agent that implements two loops: an outer loop and an inner loop. The outer loop (lighter background with solid arrows) manages the task ledger (containing facts, guesses, and plan). The inner loop (darker background with dotted arrows) manages the progress ledger (containing current progress, task assignment to agents).

- [Zhao Siyun et al. (2024)] - **Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely**

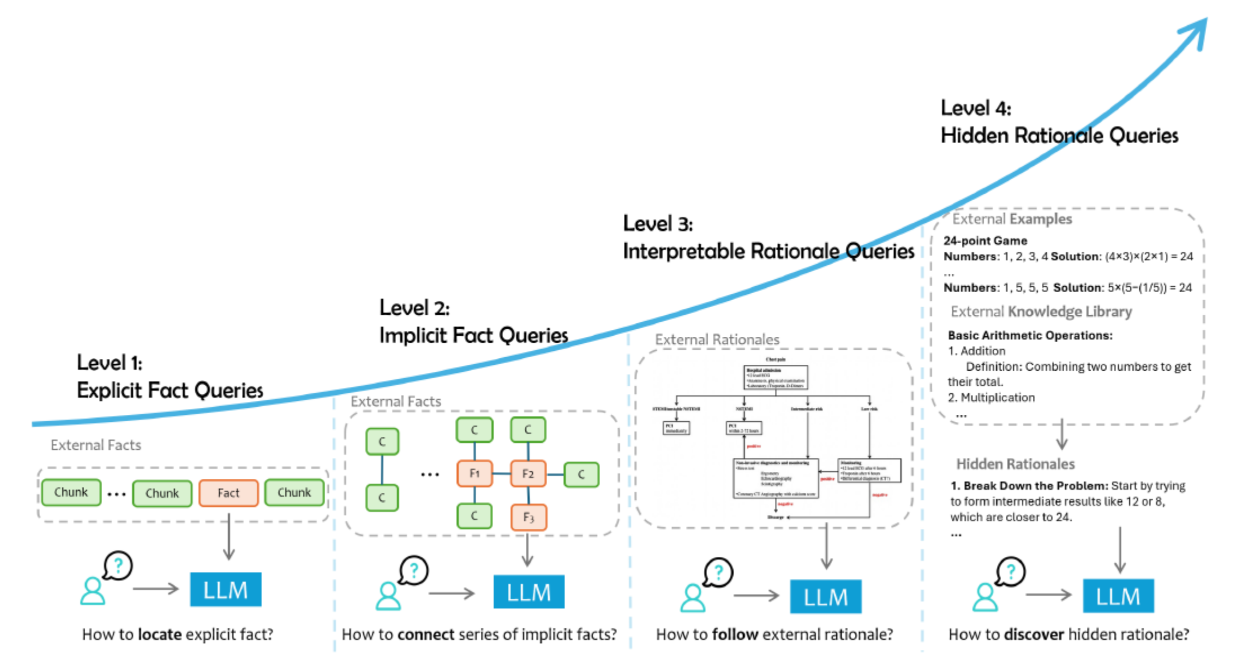

Level 1:
Explicit Fact Queries
Level 2:
Implicit Fact Queries
Level 3:
Interpretable Rationale Queries
Level 4:
Hidden Rationale Queries
External Facts
Chunk
Chunk
Fact
Chunk
External Facts
C
F1
F2
F3
External Rationales
External Examples
24-point Game
Numbers: 1, 2, 3, 4 Solution: (4×3)×(2×1) = 24
...
Numbers: 1, 5, 5, 5 Solution: 5×(5−(1/5)) = 24
External Knowledge Library
Basic Arithmetic Operations:
1. Addition
Definition: Combining two numbers to get their total.
2. Multiplication
...
Hidden Rationales
1. Break Down the Problem: Start by trying to form intermediate results like 12 or 8, which are closer to 24.
...
LLM
How to locate explicit fact?
How to connect series of implicit facts?
How to follow external rationale?
How to discover hidden rationale?

Figure 1: Main Focus of Four Level Queries

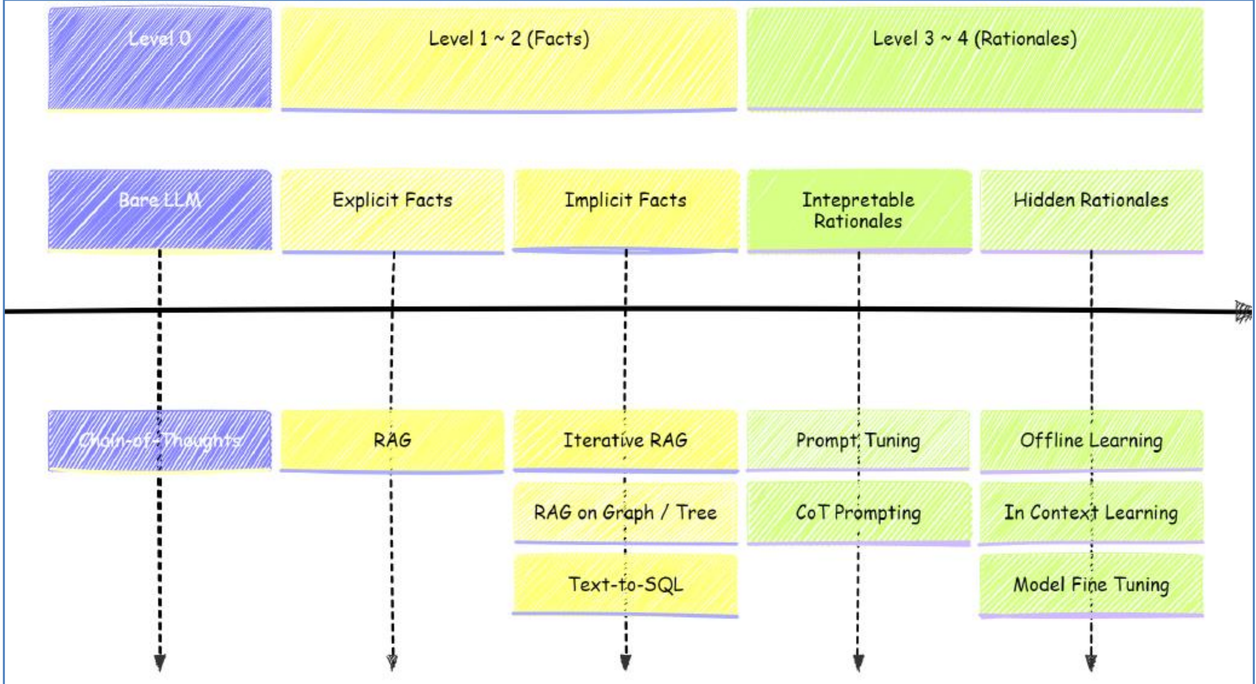


Figure 5: Summary of Main Techniques for Different Query Levels in Data augmented LLM applications

- [Pazem et al. (2024)] - **Free Energy Projective Simulation (FEPS): Active inference with interpretability**

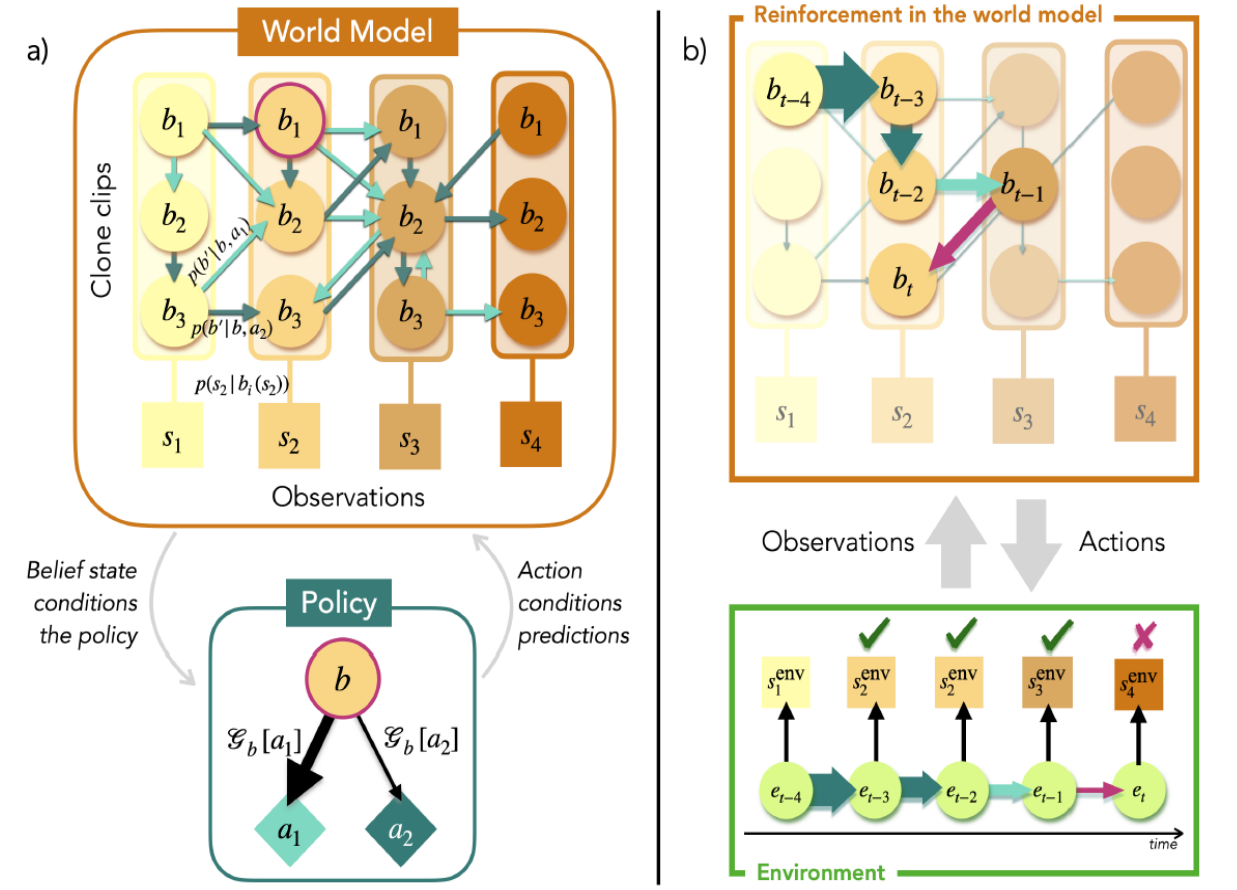


Figure 1: **Architecture and training of an FEPS agent** a) Architecture of a FEPS agent, with four sensory states (squares) and two possible actions (diamonds). The agent has two main components: the world model and the policy. The world model is composed of vertices representing observations (squares) while clone clips represent all values a belief state can take (circles). As in a clone-structured graph, each clone clip b relates to exactly one observation s and the emission function $p(s|b)$ is deterministic. The clone clips, together with the set of edges between them, form an ECM. A belief state, circled in purple, is designated by an excited clone clip. The weighted edges in the ECM encode the transition function and are trainable with reinforcement: there is one set of edges per action (light and dark turquoise arrows). The belief state in the ECM is an input to the policy, where the probability of sampling an action is a function of the EFE. In turn, the action that was selected determines the edge set to sample from in the world model in order to make a prediction for the next belief state and observation. b) Training of the world model of a FEPS agent. The agent interacts with the environment by receiving observations and implementing actions. When an action a_t is chosen, a corresponding edge $b_t \xrightarrow{a_t} b_{t+1}$ is sampled in the world model, from the current to the next belief state, conditioned on the action. The observation s_{t+1} associated with the next belief state is the prediction for the next sensory state. Simultaneously, the action is applied to the environment and creates a transition in the hidden states of the environment, $e_t \xrightarrow{a_t} e_{t+1}$ (bottom, green rectangle). This transition is perceived by the agent through the observation s_{t+1}^{env}. Finally, the weights of the edges are updated. The reinforcement of an edge is proportional to the number of correct predictions it enabled in a row, as depicted with the thickness of the arrows in the world model. When the agent makes an incorrect prediction (the purple arrow), the reinforcements are applied to the edges that contributed to the trajectory. The last, incorrect, edge is not reinforced.

- [Sumers et al. (2024)] - **Cognitive Architectures for Language Agents** *{Memory)*

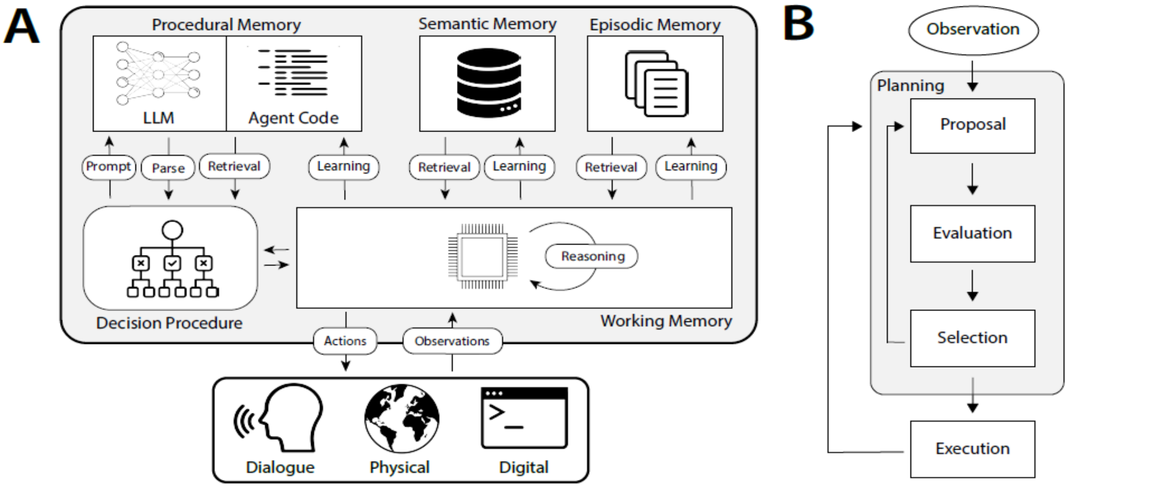


Figure 4: Cognitive architectures for language agents (CoALA). **A**: CoALA defines a set of interacting modules and processes. The **decision procedure** executes the agent's source code. This source code consists of procedures to interact with the LLM (prompt templates and parsers), internal memories (retrieval and learning), and the external environment (grounding). **B**: Temporally, the agent's decision procedure executes a **decision cycle** in a loop with the external environment. During each cycle, the agent uses **retrieval** and **reasoning** to plan by proposing and evaluating candidate **learning** or **grounding** actions. The best action is then selected and executed. An observation may be made, and the cycle begins again.

- [Park et al. (2024b)] - **Generative Agent Simulations of 1,000 People**

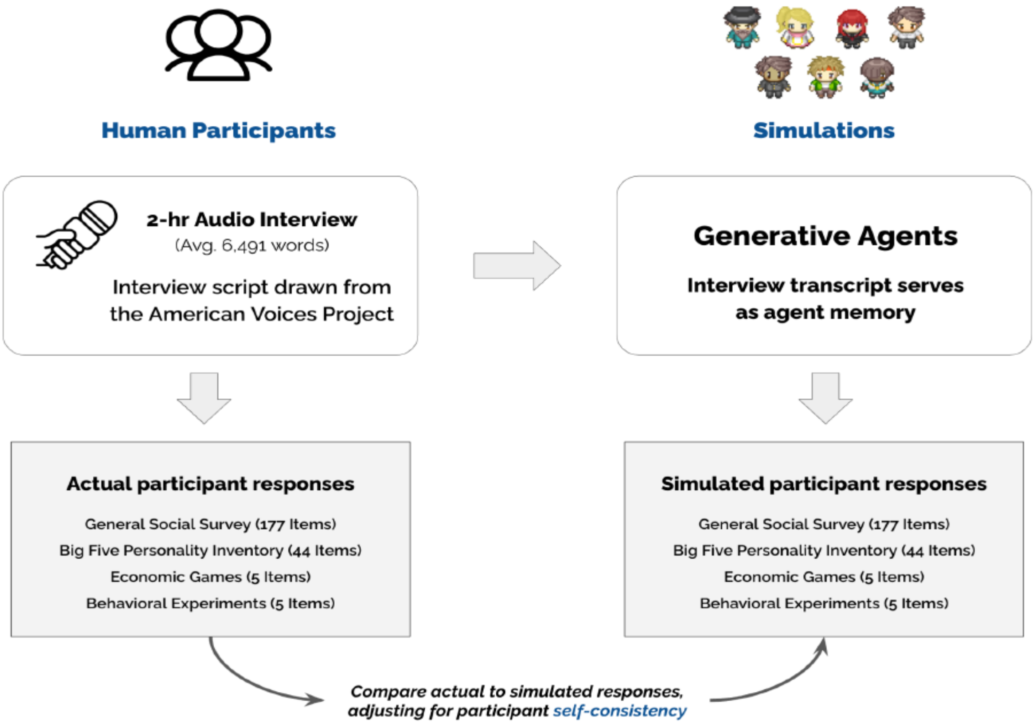

Human Participants
Simulations
2-hr Audio Interview
(Avg. 6,491 words)
Interview script drawn from the American Voices Project
Generative Agents
Interview transcript serves as agent memory
Actual participant responses
General Social Survey (177 Items)
Big Five Personality Inventory (44 Items)
Economic Games (5 Items)
Behavioral Experiments (5 Items)
Simulated participant responses
General Social Survey (177 Items)
Big Five Personality Inventory (44 Items)
Economic Games (5 Items)
Behavioral Experiments (5 Items)
Compare actual to simulated responses, adjusting for participant self-consistency

Figure 1. The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052 individuals from the U.S., selected based on age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual identity. Once recruited, participants complete a two-hour audio interview with our AI interviewer, followed by surveys and experiments. We create generative agents for each participant using their interview data. To evaluate these agents, both the generative agents and participants complete the same surveys and experiments. For the human participants, this involves retaking the surveys and experiments again two weeks later. We assess the accuracy of the agents by comparing agent responses to the participants' original responses, normalizing by how consistently each participant successfully replicates their own responses two weeks later.

- [Smirnov, Ponomarev and Agafonov (2024)] - **Ontology-Based Neuro-Symbolic AI: Effects on Prediction Quality and Explainability**

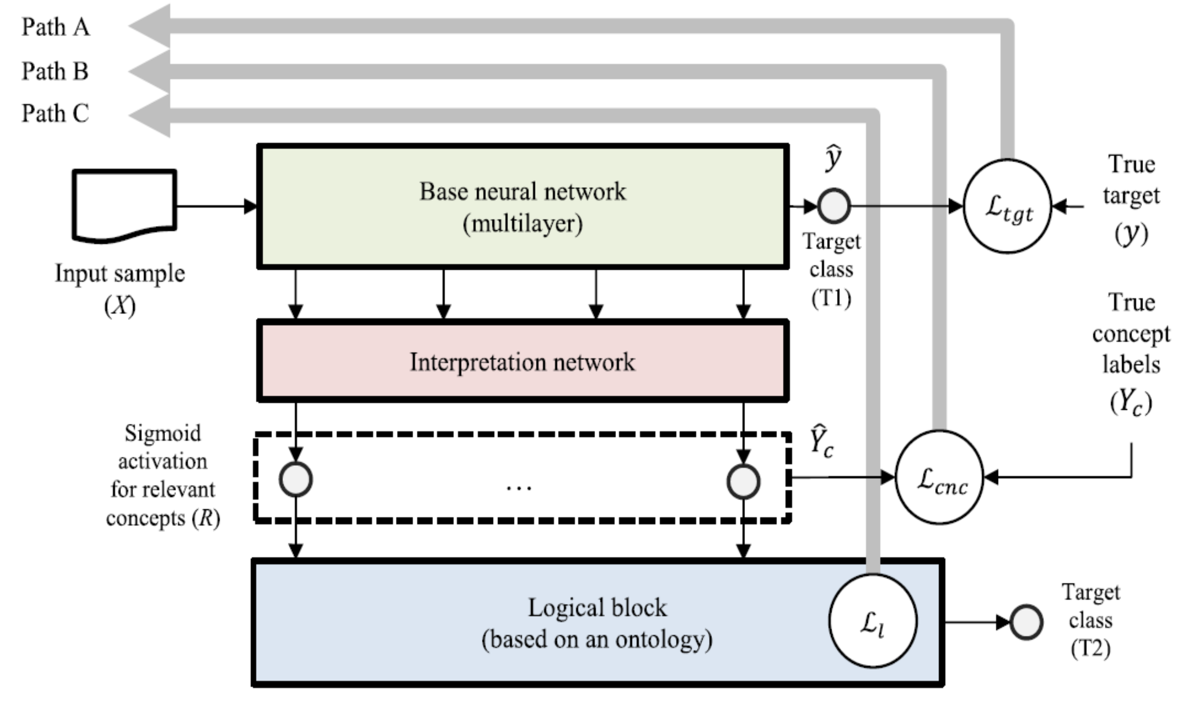


FIGURE 1. General framework for constructing explainable ontology-based neural networks.

- [Xiong et al. (2024b)] - **Converging Paradigms: The Synergy of Symbolic and Connectionist AI in LLM-Empowered Autonomous Agents** *{MAS)*

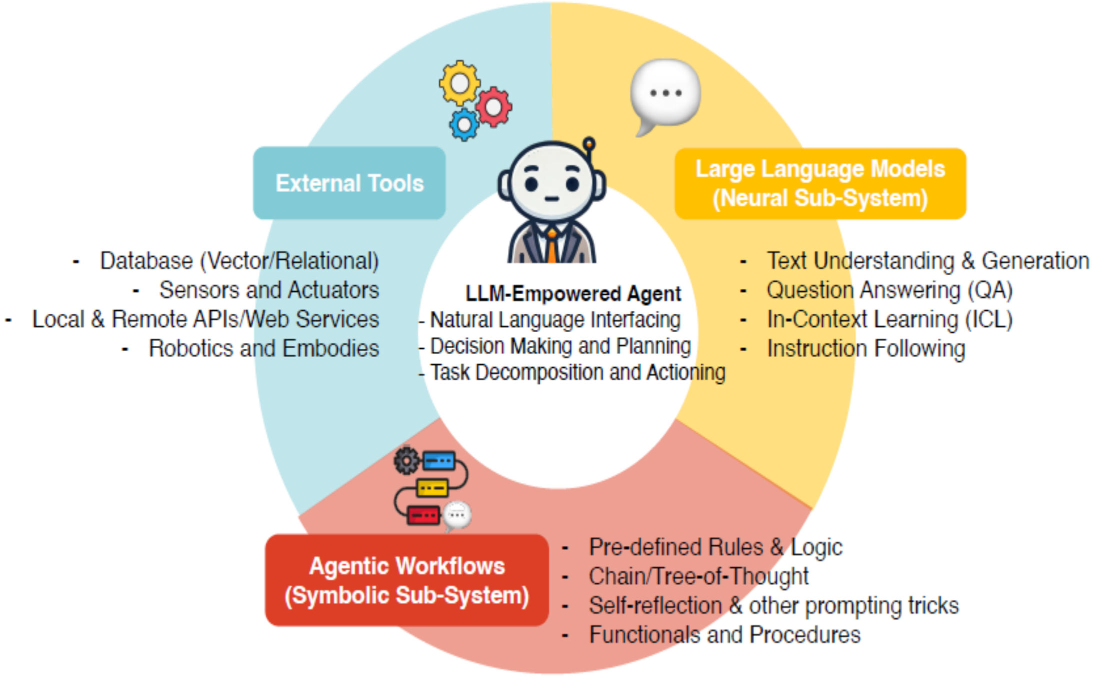

External Tools
- Database (Vector/Relational)
- Sensors and Actuators
- Local & Remote APIs/Web Services
- Robotics and Embodies
Large Language Models
(Neural Sub-System)
- Text Understanding & Generation
- Question Answering (QA)
- In-Context Learning (ICL)
- Instruction Following
LLM-Empowered Agent
- Natural Language Interfacing
- Decision Making and Planning
- Task Decomposition and Actioning
Agentic Workflows
(Symbolic Sub-System)
- Pre-defined Rules & Logic
- Chain/Tree-of-Thought
- Self-reflection & other prompting tricks
- Functionals and Procedures

Fig. 1 Elements of LLM-empowered Autonomous Agents (LAAs): Large Language Models (Neural Sub-System), Agentic Workflows (Symbolic Sub-System), and External Tools

- [Vilas et al. (2024)] **An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience**

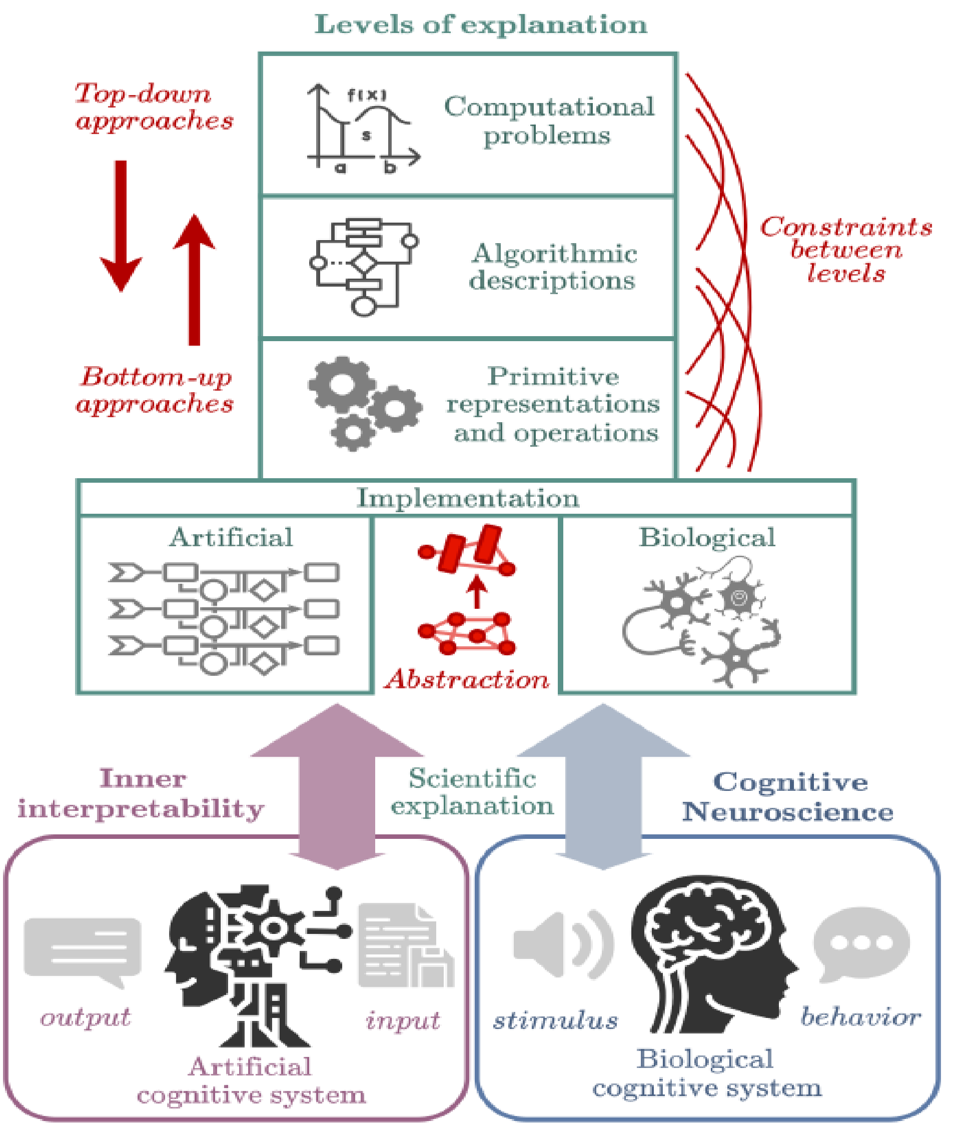


Figure 1. The fields of Inner Interpretability and Cognitive Neuroscience aim to mechanistically explain the behavior of artificial and biological systems, respectively. The multilevel explanatory framework proposed here draws out the parallels and suggests strategies that can be transferred between fields to tackle current issues in Inner Interpretability (shown in red).

- [Meta (2024)] **Large Concept Models: Language Modeling in a Sentence Representation Space** *{LLM}*

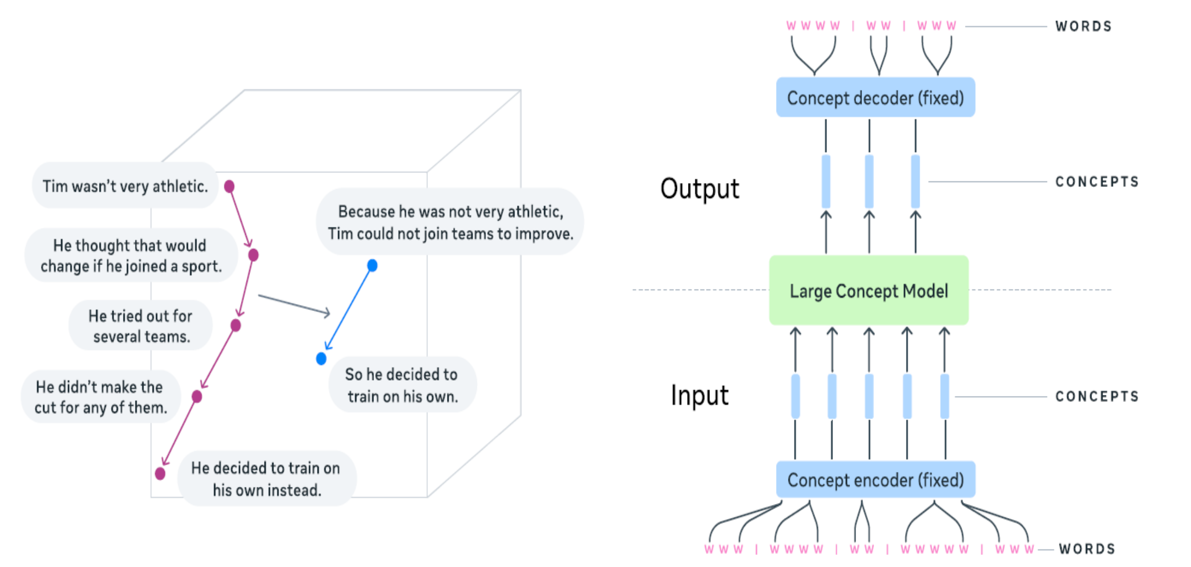

Tim wasn't very athletic.
He thought that would change if he joined a sport.
He tried out for several teams.
He didn't make the cut for any of them.
He decided to train on his own instead.
Because he was not very athletic, Tim could not join teams to improve.
So he decided to train on his own.
W W W W | W W | W W W
WORDS
Concept decoder (fixed)
Output
CONCEPTS
Large Concept Model
Input
CONCEPTS
Concept encoder (fixed)
W W W | W W W W | W W | W W W W W | W W W
WORDS

Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an Large Concept Model (LCM). $\star$: concept encoder and decoder are frozen.

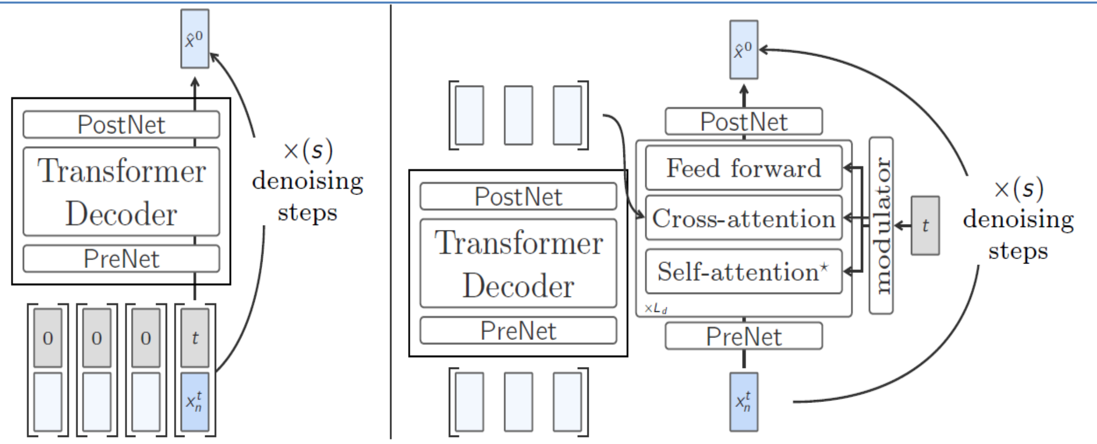


Figure 6 - Inference with diffusion-based LCMs. In the left-hand side, an illustration of the One-Tower LCM and on the right-hand side an illustration of the Two-Tower LCM.

- [Chen et al. (2024)] **Reverse Thinking Makes LLMs Stronger Reasoners**

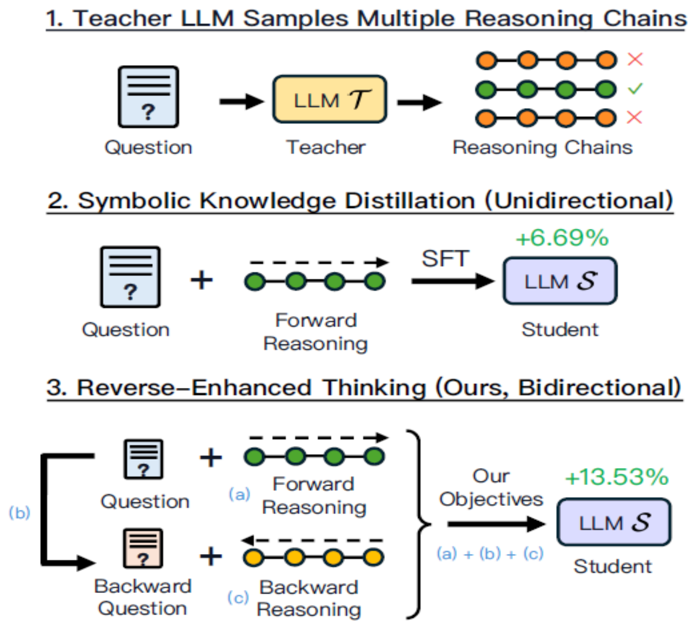


Figure 1: Comparison between symbolic knowledge distillation (SKD) and our method. (1) the teacher model generates multiple reasoning chains for a given question, (2) SKD supervised fine-tunes on the correct reasoning chains, and (3) our method incorporates bidirectional reasoning, learning from both Q-to-A and A-to-Q using our multi-task objectives.

- [Waade et al. (2024)] **As One and Many: Relating Individual and Emergent Group-Level Generative Models in Active Inference**

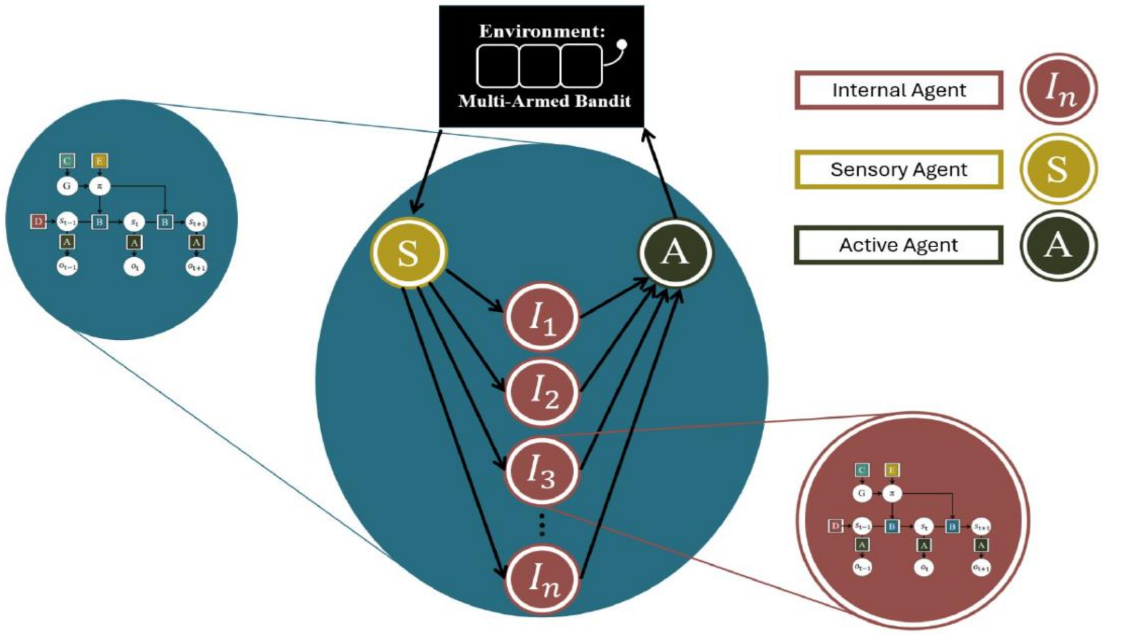


Figure 3. The agent group structure

[Bhaskar & Kuppan (2024)] - **Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive Survey** *{Agent, Architecture, Learning, Training}*

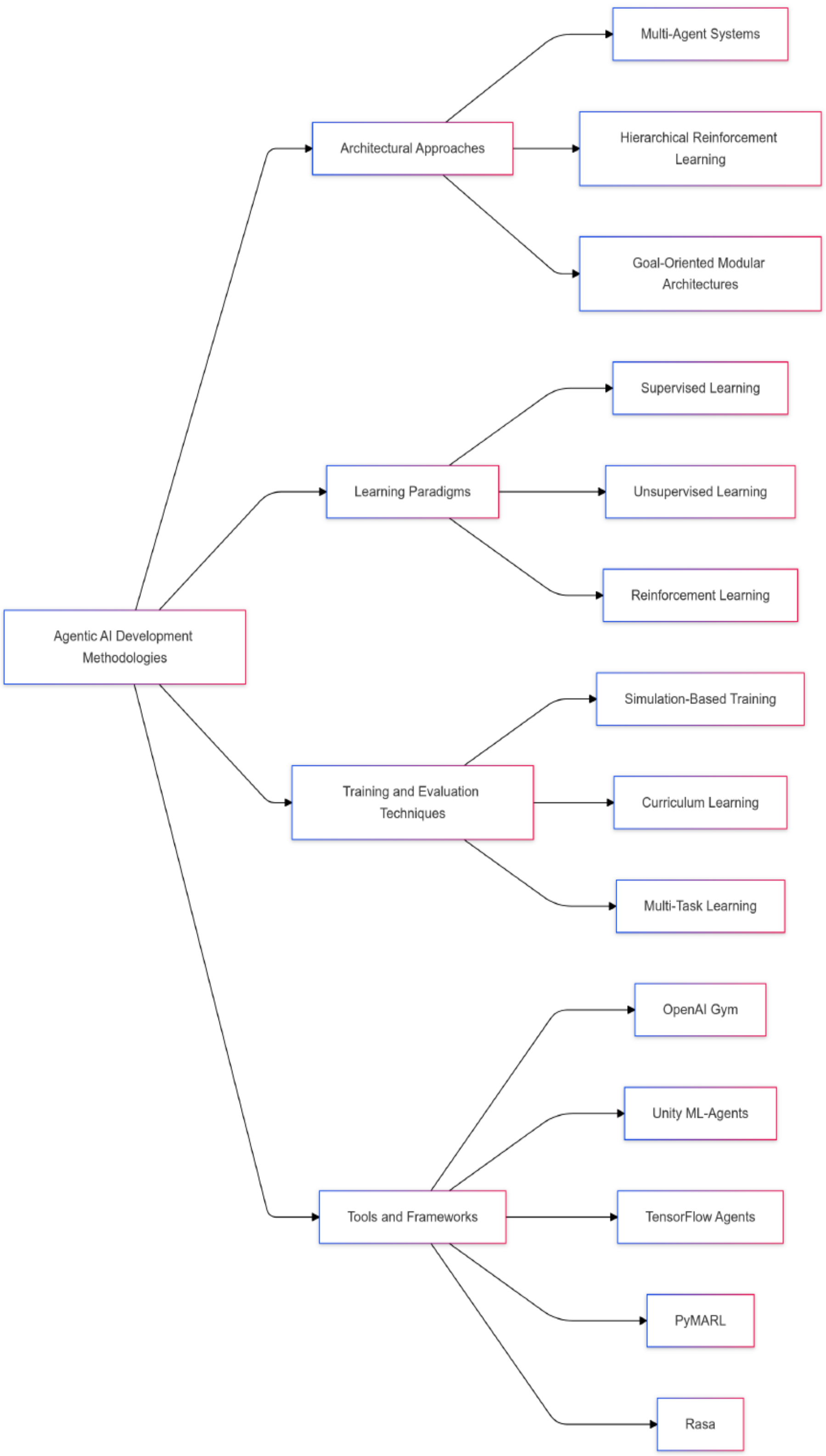


FIGURE 3: Overview of Agentic AI Development Methodologies, including Architectural Approaches, Learning Paradigms, Training Techniques, and Tools.

- [Barnes & Hutson (2024)] - **AI and the Cognitive Sense of Self**

Table 1. Key Components and Implementations of Cognitive Sense of Self in AI Systems

Component	Definition	Implementation
Self-Recognition	Ability of an AI system to identify itself as distinct from its environment and other entities.	Techniques such as computer vision and proprioception are utilized to help AI systems discern their physical presence and distinguish themselves from external objects.
Self-Reflection	Capacity of an AI system to monitor and evaluate its own internal states, processes, and behaviors.	AI systems maintain logs of their actions and outcomes, analyze this data to detect patterns, and adjust their strategies accordingly. Machine learning algorithms play a critical role in enabling the system to learn from past experiences.
Continuity of Identity	Involves maintaining a consistent sense of self over time.	Memory systems and data storage preserve information about past states and actions, allowing AI systems to build a coherent narrative of their existence. Techniques such as long-term memory in neural networks and temporal coherence algorithms support this continuity.
Agency and Intentionality	Refers to the AI system's ability to act upon its environment based on internal goals and motivations.	AI systems are designed with goal-setting mechanisms and motivational frameworks that drive their behavior. Reinforcement learning algorithms help AI agents develop strategies to achieve their goals based on rewards and feedback from the environment.
Self-Monitoring and Error Correction	Ongoing process of checking and evaluating one's own performance and rectifying mistakes.	Diagnostic tools and self-repair mechanisms are integrated into AI systems for continuous self-monitoring and error correction. Machine learning models that predict and detect anomalies assist systems in identifying errors in real-time and taking corrective actions.
Enhanced Decision-Making and Autonomy	Allows AI agents to make autonomous and well-informed decisions based on their state and capabilities.	AI systems can evaluate options and choose actions that align with their goals and constraints, especially important in dynamic and unpredictable environments where pre-programed responses are insufficient.
Adaptive Learning and Behavior	AI systems benefit from the ability to reflect on past actions and outcomes to enhance performance.	By learning from experiences and adapting over time, AI systems can continually optimize their performance, crucial for long-term deployment and continuous improvement.
Meaningful Human-AI Interaction	AI agents can achieve more intuitive and natural interactions with humans.	AI systems understand and respond to human social cues, anticipate needs, and provide personalized assistance, essential for applications in customer service, healthcare, and collaborative robotics.

Table 2. Mechanisms and Roles in Developing Identity in Artificial Intelligence Systems

Aspect	Definition	Details and Citations
Memory in Identity Development	Crucial for maintaining a continuous sense of identity.	Continuity of Experience: Enables AI to store and retrieve past states, actions, and experiences to construct a coherent narrative of their existence Contextual Awareness: Helps AI make informed decisions by applying lessons learned from past experiences to new situations, enhancing adaptability and depth of identity
Learning in Identity Development	Central to the evolution of AI identity through adaptation and personalization.	Adaptive Behavior: Allows AI to modify and improve actions based on new information and experiences, driven by machine learning algorithms such as reinforcement learning and neural network training Personalized Growth: Supports development of unique characteristics by tailoring learning processes to specific interactions and experiences
Self-Recognition in Identity Development	Enables AI to distinguish itself from its environment and other agents, fostering autonomy and self-awareness.	Physical and Functional Self-Recognition: Technologies such as computer vision and proprioception allow AI to recognize its own physical form and movements, essential for distinguishing self-generated actions from external events Internal State Monitoring: Enhances self-recognition by monitoring internal states and processes, aiding in maintaining a consistent self-image and adapting behaviors

- [Zhang & Xu (2024)] **- An Overview of the Free Energy Principle and Related Research**

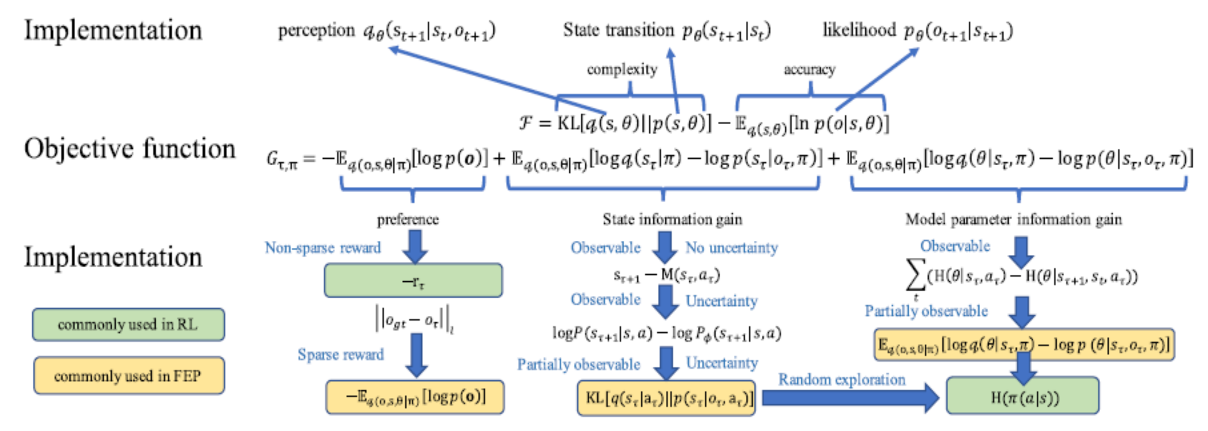

Implementation
perception $q_\theta(s_{t+1}|s_t, o_{t+1})$
State transition $p_\theta(s_{t+1}|s_t)$
likelihood $p_\theta(o_{t+1}|s_{t+1})$
complexity
accuracy
Objective function
$\mathcal{F} = \mathrm{KL}[q(s,\theta)||p(s,\theta)] - \mathbb{E}_{q(s,\theta)}[\ln p(o|s,\theta)]$
$G_{\tau,\pi} = -\mathbb{E}_{q(o,s,\theta|\pi)}[\log p(\boldsymbol{o})] + \mathbb{E}_{q(o,s,\theta|\pi)}[\log q(s_\tau|\pi) - \log p(s_\tau|o_\tau,\pi)] + \mathbb{E}_{q(o,s,\theta|\pi)}[\log q(\theta|s_\tau,\pi) - \log p(\theta|s_\tau,o_\tau,\pi)]$
preference
State information gain
Model parameter information gain
Implementation
Non-sparse reward
$-r_\tau$
$||o_{gt} - o_\tau||_l$
Sparse reward
$-\mathbb{E}_{q(o,s,\theta|\pi)}[\log p(\mathbf{o})]$
Observable
No uncertainty
$s_{\tau+1} - \mathrm{M}(s_\tau, a_\tau)$
Observable
Uncertainty
$\log P(s_{\tau+1}|s,a) - \log P_\phi(s_{\tau+1}|s,a)$
Partially observable
Uncertainty
$\mathrm{KL}[q(s_\tau|a_\tau)||p(s_\tau|o_\tau,a_\tau)]$
Random exploration
Observable
$\sum_t (\mathrm{H}(\theta|s_\tau,a_\tau) - \mathrm{H}(\theta|s_{\tau+1},s_t,a_\tau))$
Partially observable
$\mathbb{E}_{q(o,s,\theta|\pi)}[\log q(\theta|s_\tau,\pi) - \log p(\theta|s_\tau,o_\tau,\pi)]$
$\mathrm{H}(\pi(a|s))$
commonly used in RL
commonly used in FEP

Figure 2: Free energy and the objective of model-based RL.

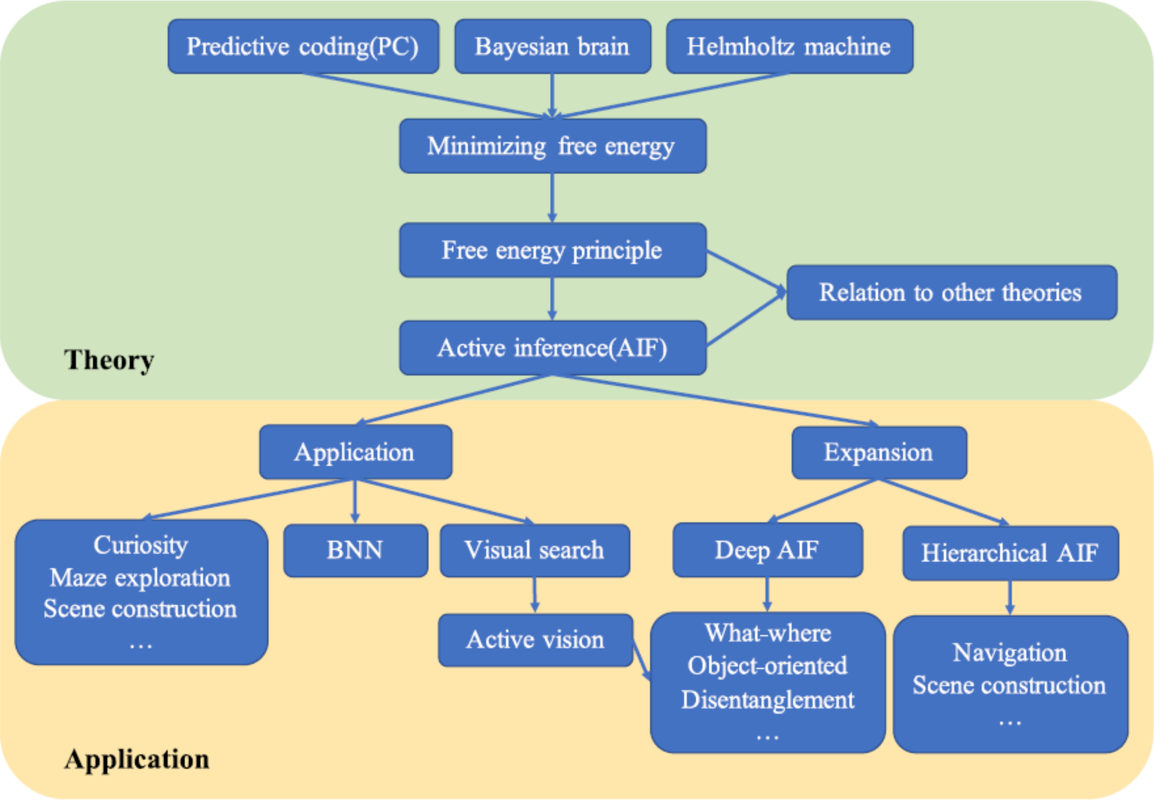


Figure 4: Developmental history of the free energy principle (FEP).

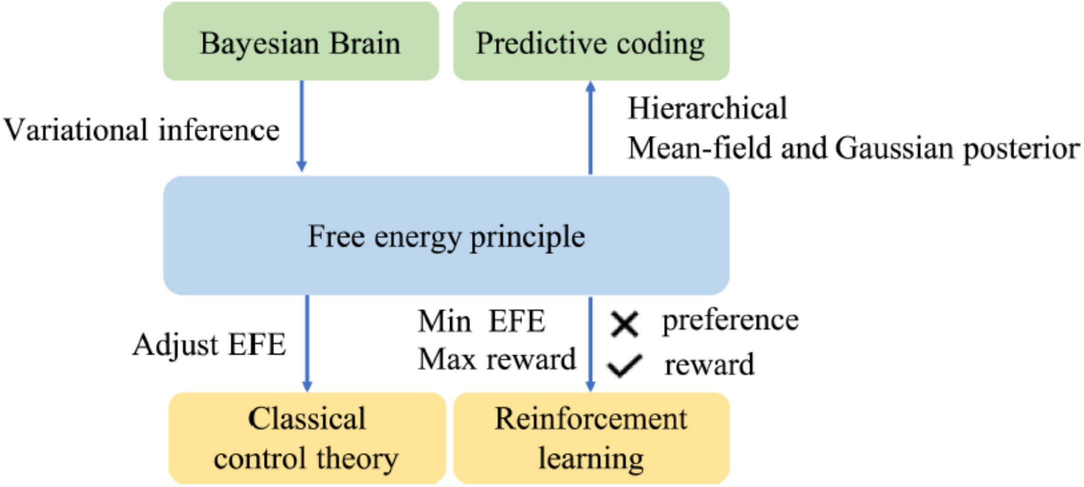


Figure 5: Relation to other theories.

{Free Energy Principle, Active Inference, Bayesian, RL}

- [Kornieiev (2025)] - **Cognitive architecture AGICA: "Space of Reasoning of individual common sense"** *{AGI, Consciousness, functions}*

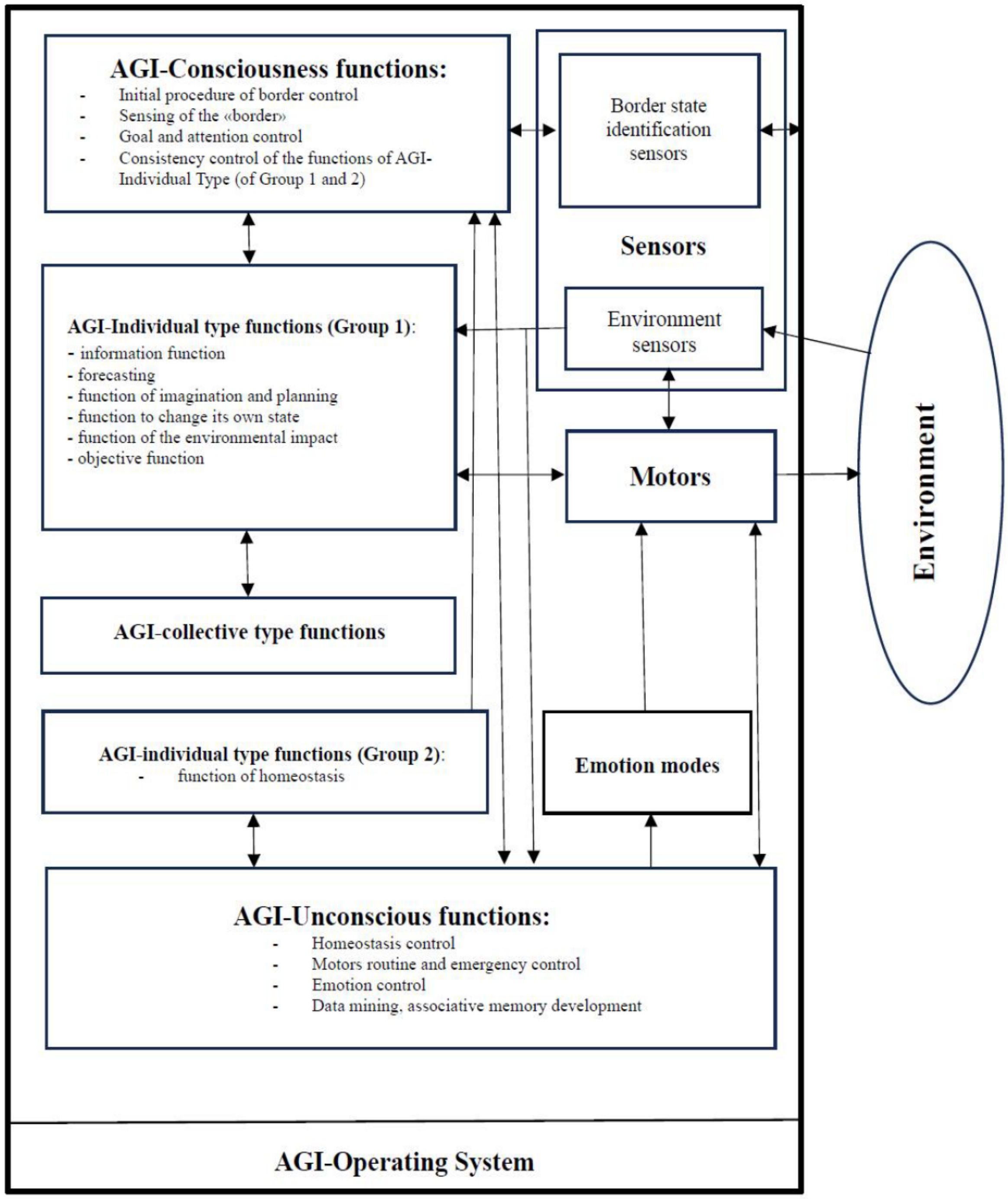

AGI-Consciousness functions:
- Initial procedure of border control
- Sensing of the «border»
- Goal and attention control
- Consistency control of the functions of AGI-Individual Type (of Group 1 and 2)
Border state identification sensors
Sensors
Environment sensors
AGI-Individual type functions (Group 1):
- information function
- forecasting
- function of imagination and planning
- function to change its own state
- function of the environmental impact
- objective function
Motors
Environment
AGI-collective type functions
AGI-individual type functions (Group 2):
- function of homeostasis
Emotion modes
AGI-Unconscious functions:
- Homeostasis control
- Motors routine and emergency control
- Emotion control
- Data mining, associative memory development
AGI-Operating System

Fig.1 Cognitive Architecture AGICA - functional view.

R. How Far Are We From AGI?

We will devote this chapter to well-structured and systematic paper about AGI – **[Feng et al. (2024)]** Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, Jiaxuan You. How Far Are We From AGI? arXiv:2405.10313v1 [cs.AI] 16 May 2024. Contents and some key illustrations here:

1. Introduction

2. AGI Internal: Unveiling the Mind of AGI

2.1. AI Perception

2.2. AI Reasoning

2.3. AI Memory

2.4. 2.4 AI Metacognition

3. AGI Interface: Connecting the World with AGI

3.1. AI Interfaces to Digital World

3.2. AI Interfaces to Physical World

3.3. AI Interfaces to Intelligence

3.3.1.AI Interface to Other AI agents

3.3.2.AI Interfaces to Humans

4. AGI Systems: Implementing the Mechanism of AGI

4.1. System Challenges

4.2. Scalable Model Architectures

4.3. Large-scale Training

4.4. Inference Techniques

4.5. Cost and Efficiency

4.6. Computing Platforms

4.7. The Future of AGI Systems

5. AGI Alignment: Ensuring AGI Meets Various Needs

5.1. Expectations of AGI Alignment

5.2. Current Alignment Techniques

5.3. How to approach AGI Alignments

6. AGI Roadmap: Responsibly Approaching AGI

6.1. AGI Levels

6.2. AGI Evaluation

6.2.1.Expectations for AGI Evaluation

6.2.2.Current Evaluations and Their Limitations

6.3. How to Get to the Next AGI Level

6.4. “How Far Are We from AGI” Workshop Discussions

6.5. Further Considerations during AGI Development

7. Case Studies: A Bright Future with AGI

7.1. AI for Science Discovery and Research

7.2. Generative Visual Intelligence

7.3. World Models for AGI

7.4. Decentralized AI

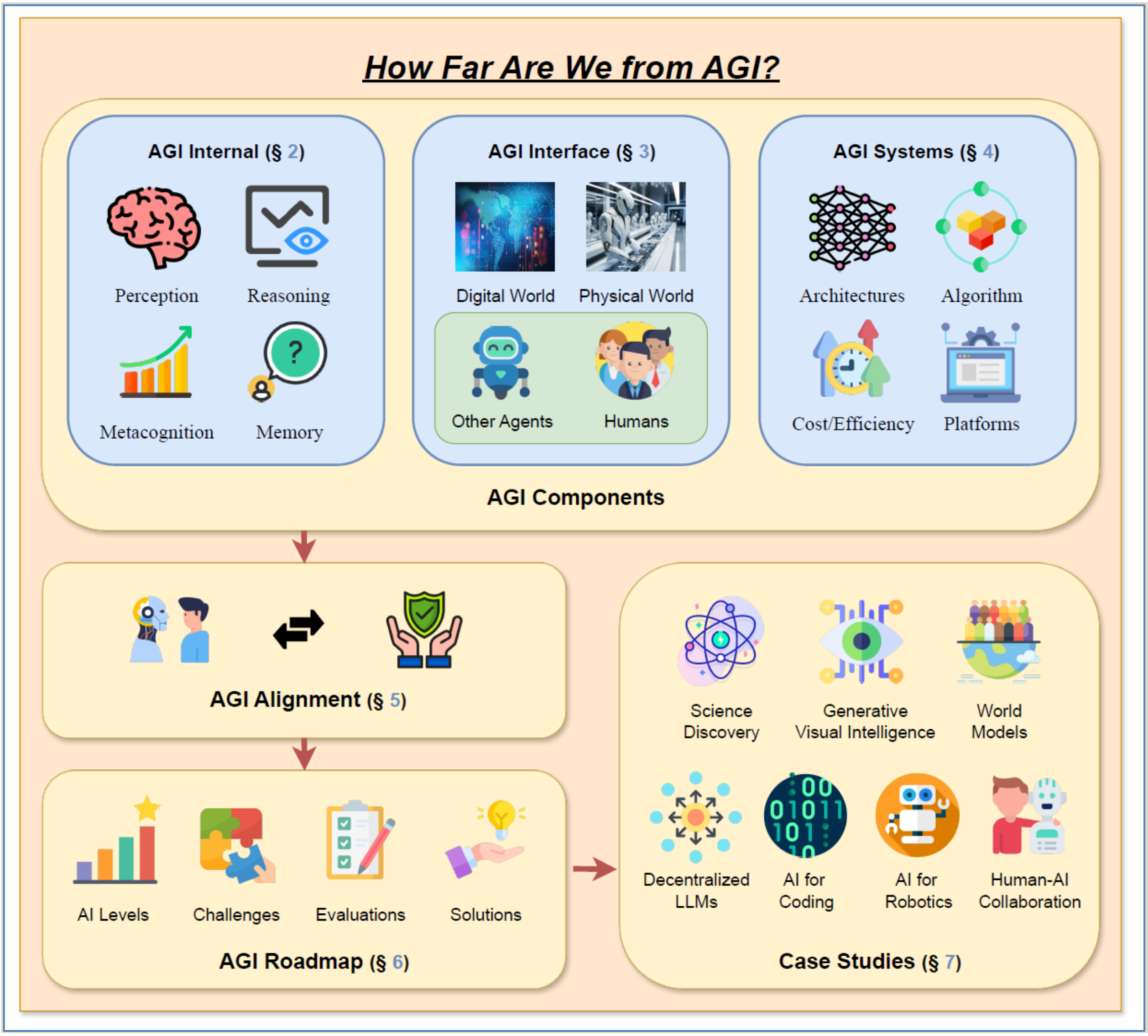
7.5. AI for Coding

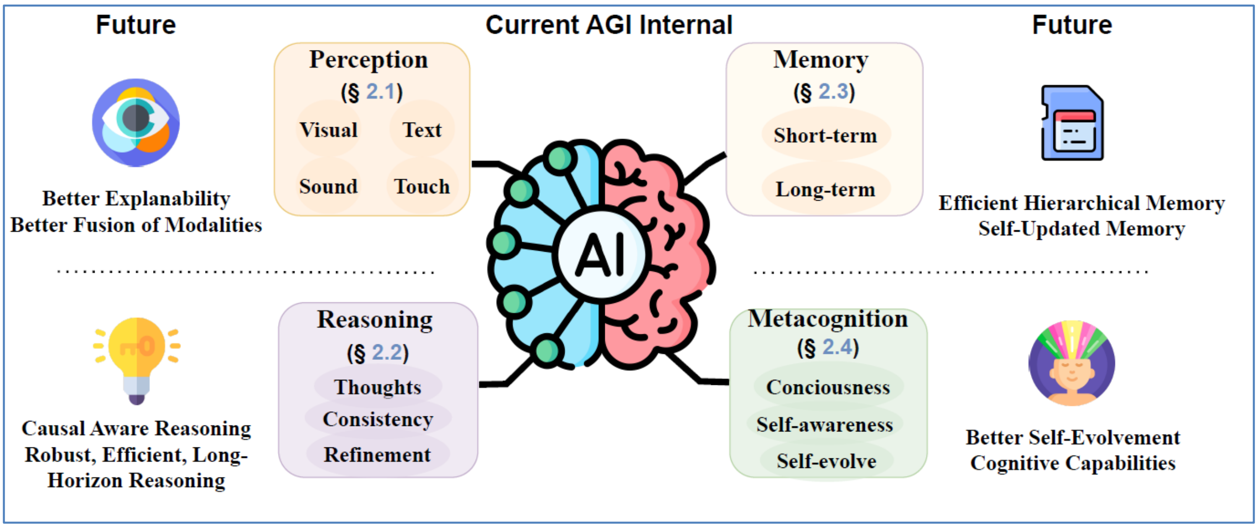
7.6. Embodied AI: AI for Robotics

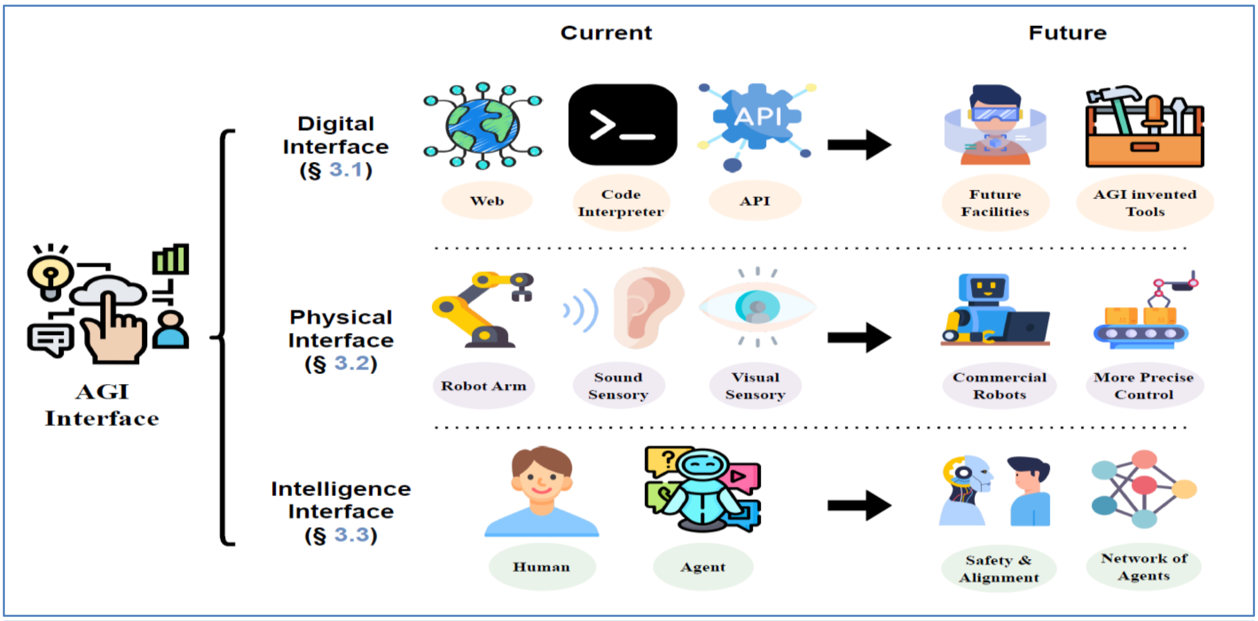
7.7. Human-AI Collaboration

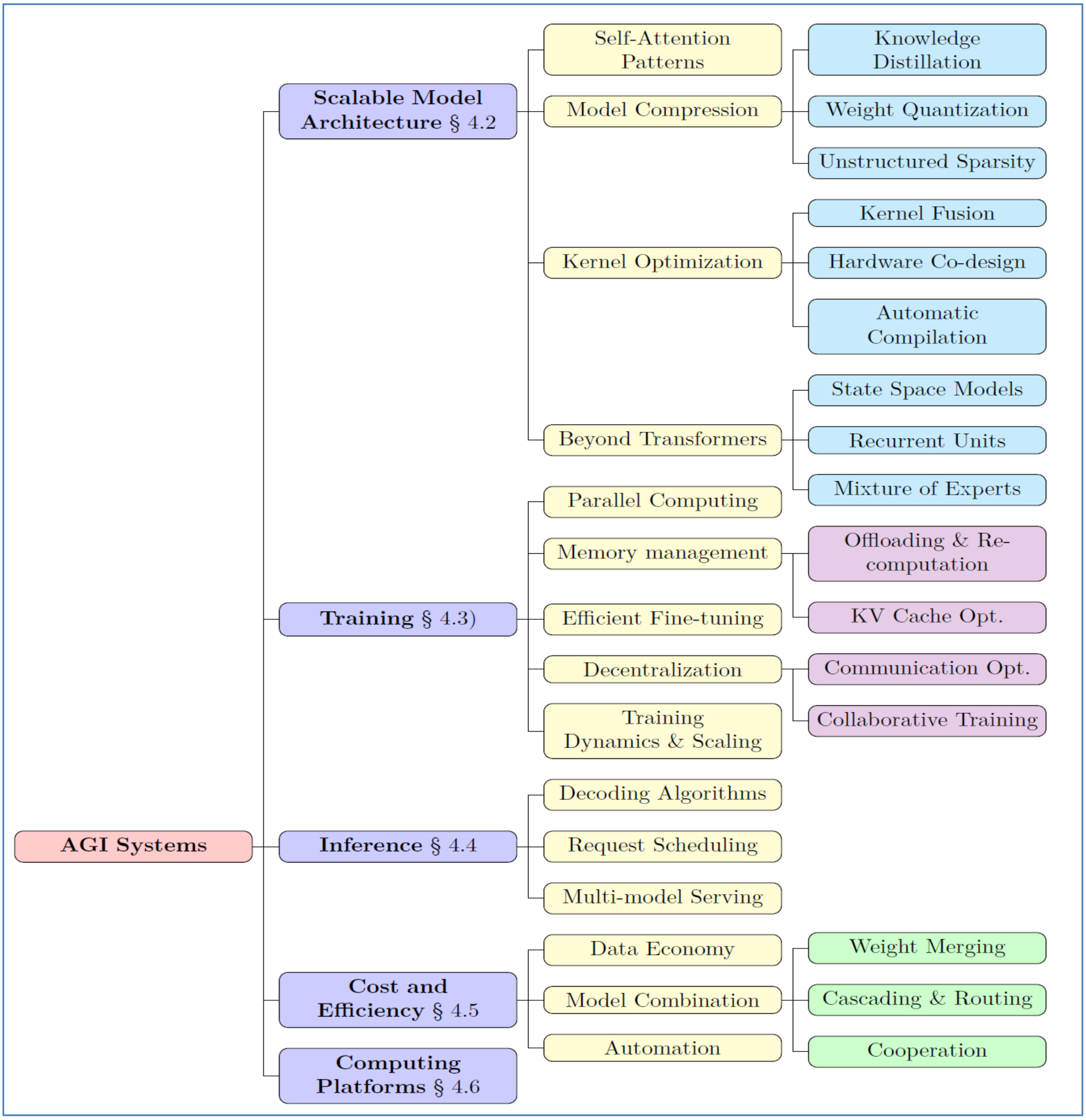
Two key references for this Paper (all reference list total 40+ pages!!):

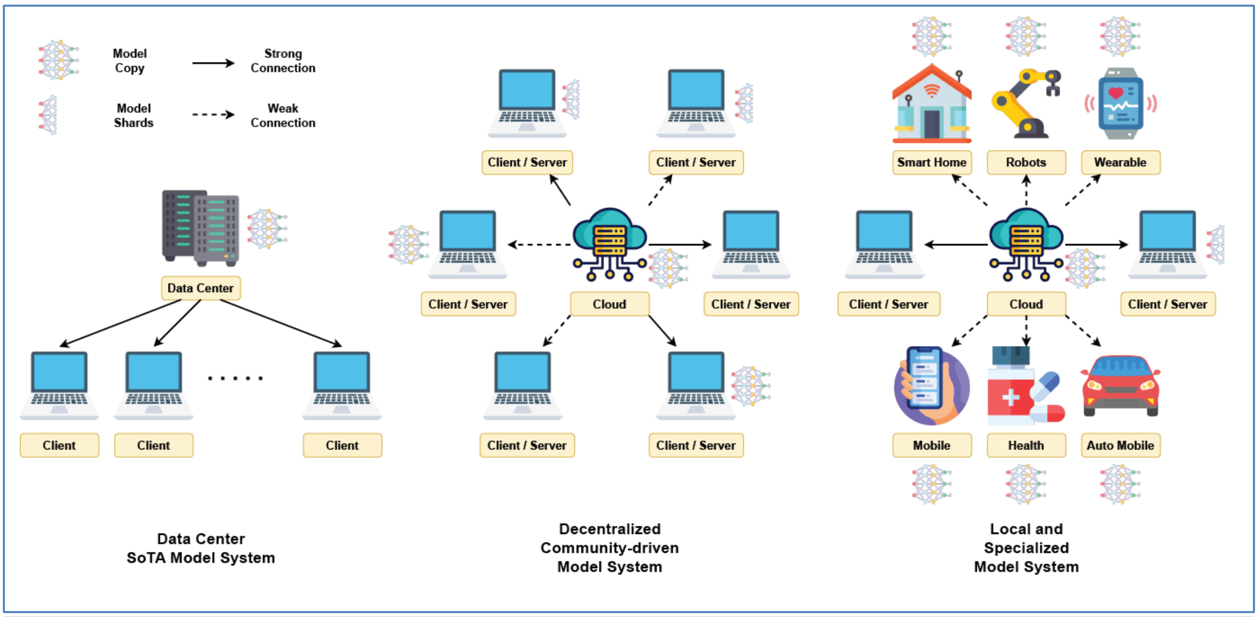
- [Voss & Jovanovic (2023a)] - **Concepts is All You Need: A More Direct Path to AGI**
- [Morris et al. (2023)] - **Levels of AGI: Operationalizing Progress on the Path to AGI**

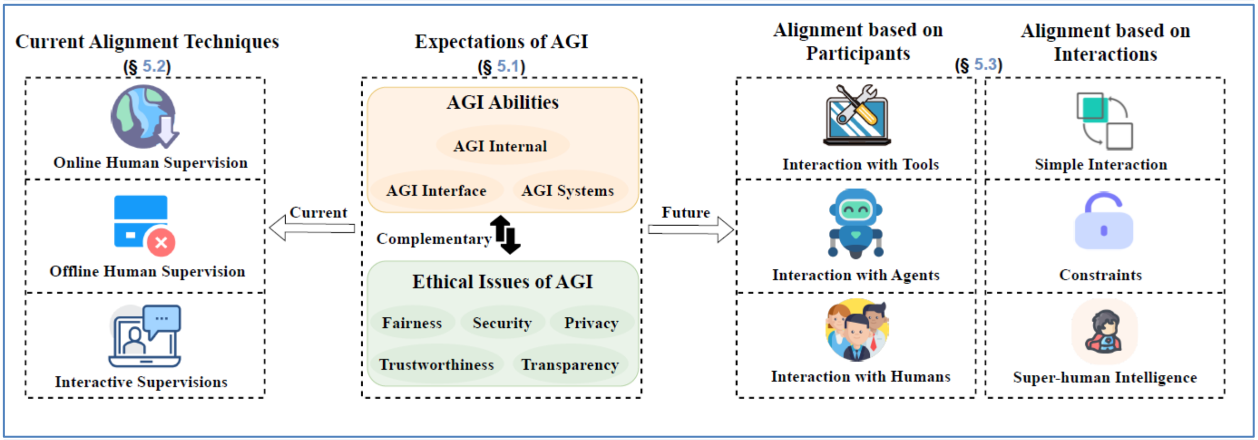


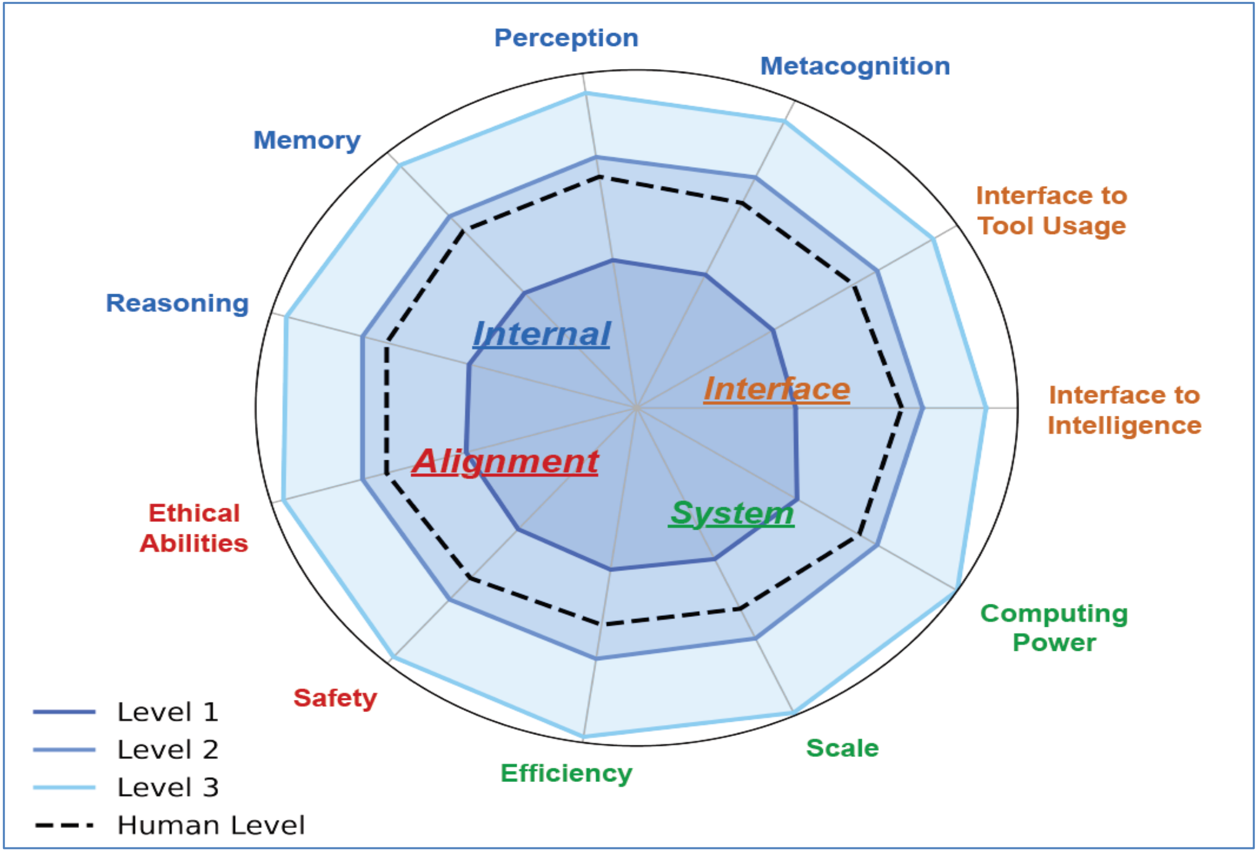


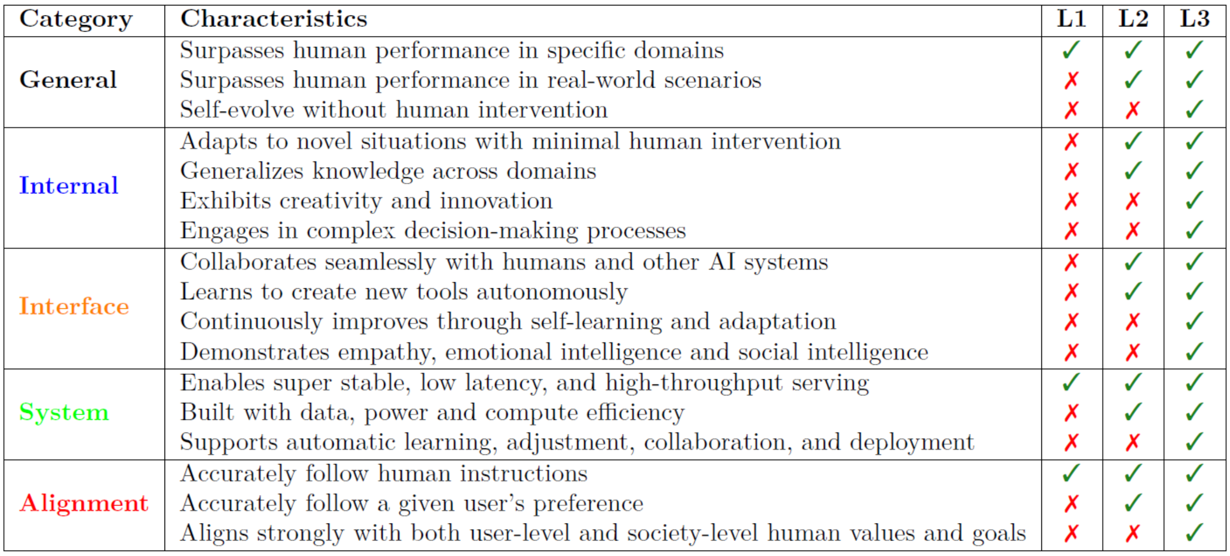

Current
Future
Digital Interface (§ 3.1)
Web
Code Interpreter
API
Future Facilities
AGI invented Tools
Physical Interface (§ 3.2)
Robot Arm
Sound Sensory
Visual Sensory
Commercial Robots
More Precise Control
Intelligence Interface (§ 3.3)
Human
Agent
Safety & Alignment
Network of Agents
AGI Interface


AGI Systems
Scalable Model Architecture § 4.2
Self-Attention Patterns
Model Compression
Knowledge Distillation
Weight Quantization
Unstructured Sparsity
Kernel Optimization
Kernel Fusion
Hardware Co-design
Automatic Compilation
Beyond Transformers
State Space Models
Recurrent Units
Mixture of Experts
Training § 4.3)
Parallel Computing
Memory management
Offloading & Re-computation
KV Cache Opt.
Efficient Fine-tuning
Decentralization
Communication Opt.
Collaborative Training
Training Dynamics & Scaling
Inference § 4.4
Decoding Algorithms
Request Scheduling
Multi-model Serving
Cost and Efficiency § 4.5
Data Economy
Model Combination
Automation
Weight Merging
Cascading & Routing
Cooperation
Computing Platforms § 4.6


Model Copy
Strong Connection
Model Shards
Weak Connection
Data Center
Client
Client
Client
Data Center
SoTA Model System
Client / Server
Client / Server
Client / Server
Cloud
Client / Server
Client / Server
Client / Server
Decentralized
Community-driven
Model System
Smart Home
Robots
Wearable
Client / Server
Cloud
Client / Server
Mobile
Health
Auto Mobile
Local and
Specialized
Model System


Current Alignment Techniques
(§ 5.2)
Online Human Supervision
Offline Human Supervision
Interactive Supervisions
Current
Expectations of AGI
(§ 5.1)
AGI Abilities
AGI Internal
AGI Interface
AGI Systems
Complementary
Ethical Issues of AGI
Fairness
Security
Privacy
Trustworthiness
Transparency
Future
Alignment based on Participants
(§ 5.3)
Alignment based on Interactions
Interaction with Tools
Interaction with Agents
Interaction with Humans
Simple Interaction
Constraints
Super-human Intelligence


Perception
Metacognition
Memory
Interface to Tool Usage
Reasoning
Interface to Intelligence
Internal
Interface
Alignment
System
Ethical Abilities
Computing Power
Safety
Scale
Efficiency
Level 1
Level 2
Level 3
Human Level



Category	Characteristics	L1	L2	L3
General	Surpasses human performance in specific domains	✓	✓	✓
	Surpasses human performance in real-world scenarios	✗	✓	✓
	Self-evolve without human intervention	✗	✗	✓
Internal	Adapts to novel situations with minimal human intervention	✗	✓	✓
	Generalizes knowledge across domains	✗	✓	✓
	Exhibits creativity and innovation	✗	✗	✓
	Engages in complex decision-making processes	✗	✗	✓
Interface	Collaborates seamlessly with humans and other AI systems	✗	✓	✓
	Learns to create new tools autonomously	✗	✓	✓
	Continuously improves through self-learning and adaptation	✗	✗	✓
	Demonstrates empathy, emotional intelligence and social intelligence	✗	✗	✓
System	Enables super stable, low latency, and high-throughput serving	✓	✓	✓
	Built with data, power and compute efficiency	✗	✓	✓
	Supports automatic learning, adjustment, collaboration, and deployment	✗	✗	✓
Alignment	Accurately follow human instructions	✓	✓	✓
	Accurately follow a given user's preference	✗	✓	✓
	Aligns strongly with both user-level and society-level human values and goals	✗	✗	✓

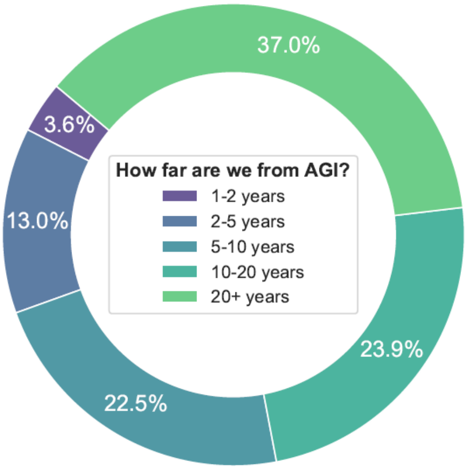


Figure 9: **Polls Results of Researchers' Opinions on When AGI Will be Achieved.** Among the attending researchers at the ICLR 2024 "How Far Are We From AGI" workshop, a survey is conducted to gather their opinions on how far they think AGI will be achieved. A total of 138 responses are received as above. Interestingly, 37% of researchers think it will take more than 20 years from now on to realize AGI.

Key findings for our Project

- AGI functions
- AGI system components
- AGI levels and characteristics
- AGI roadmap
- AGI case studies
- In general - well-structured, comprehensive and systematic Work!

S. ASI Project

Significant and promising ASI (Sic!) Project from Dr. Ben Goertzel, Team and Partners:

Key references:

- [Goertzel (2006)] - **The Hidden Pattern**
- [Goertzel, Ikle' & Wigmore (2012)] - **The architecture of human-like general intelligence**
- [Goertzel, Pennachin & Geisweiller (2013a)] - **Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy**
- [Goertzel, Pennachin & Geisweiller (2013b)] - **Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI**
- [Goertzel (2014a)] -**The AGI Revolution**
- [Goertzel (2014b)] - **Golem: towards an agi meta-architecture enabling both goal preservation and radical self-improvement**
- [Goertzel (0017a)] -**Toward a formal model of cognitive synergy**
- [Goertzel (2017b)] - **Euryphysics: a (somewhat) new conceptual model of mind, reality and psi**
- [Goertzel (2019)] - **Distinction graphs and graphtropy: A formalized phenomenological layer underlying classical and quantum entropy, observational semantics and cognitive computation**
- [Goertzel (2020)] - **Paraconsistent foundations for probabilistic reasoning, programming and concept formation**
- [Goertzel (2021b)] - **Patterns of cognition: Cognitive algorithms as galois connections fulfilled by chronomorphisms on probabilistically typed metagraphs**

The General Theory of General Intelligence

Theoretical (Cognitology) Concept based on **[Goertzel (2021a)]** Ben Goertzel. The General Theory of General Intelligence: A Pragmatic Patternist Perspective. arXiv:2103.15100v3 [cs.AI] 4 Apr 2021

1. **Introduction**
    1.1. Summary of Key Points
2. **Patternist Philosophy of Mind**
    2.1. Patternist Principles
    2.2. Cognitive Synergy
3. **Foundational Ontology**
    3.1. From Laws of Form to Paraconsistent and Probabilistic Logic
    3.2. From Distinction Graphs to Dynamic Knowledge Metagraphs
        3.2.1.Distinctions Transcending Distinctions
    3.3. Measuring Simplicity and Pattern
    3.4. Associativity and Subpattern Hierarchy
        3.4.1.From Subpattern Hierarchies to Dual Networks
    3.5. Generalized Probabilities
4. **Quantifying General Intelligence**
    4.1. General Intelligence as Expected Reward Maximization Performance
    4.2. Pragmatic General Intelligence
    4.3. Intellectual Breadth
    4.4. Multiple Criterion Driven General Intelligence
5. **Universal Algorithms for General Intelligence**

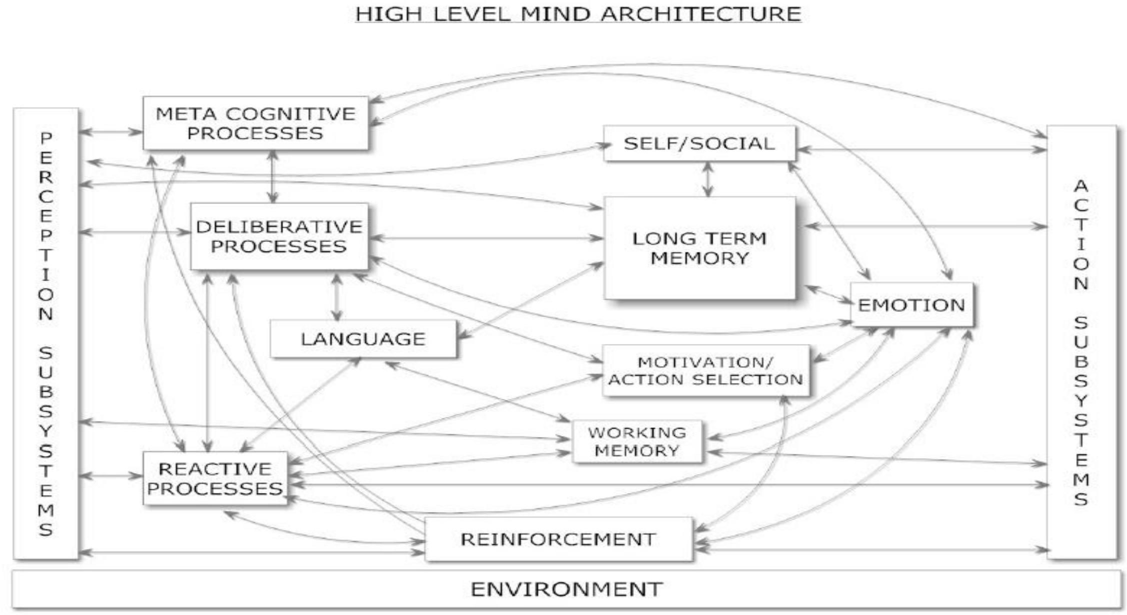


Figure 17: High level architecture diagram for a human-like general intelligence, inspired by the work of Aaron Sloman among other sources. This can be viewed in an obvious way as a particular way of refining Figure [?]. From [GIW12]

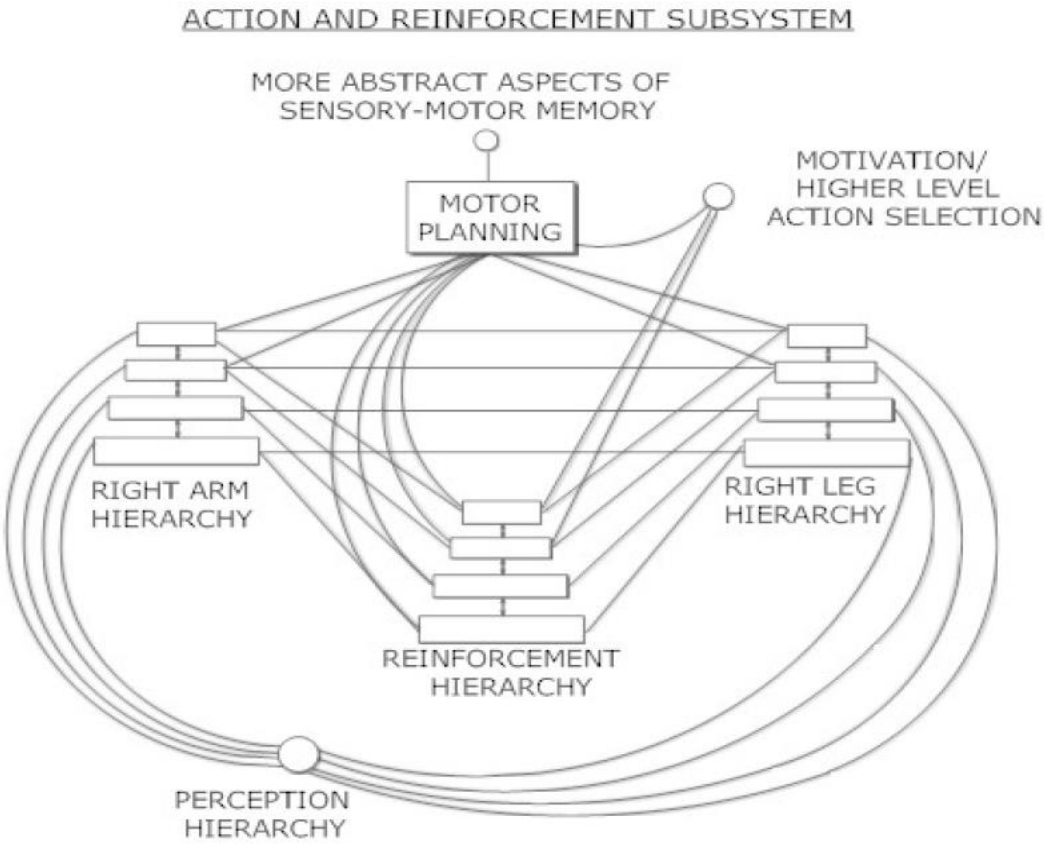

ACTION AND REINFORCEMENT SUBSYSTEM
MORE ABSTRACT ASPECTS OF SENSORY-MOTOR MEMORY
MOTOR PLANNING
MOTIVATION/ HIGHER LEVEL ACTION SELECTION
RIGHT ARM HIERARCHY
RIGHT LEG HIERARCHY
REINFORCEMENT HIERARCHY
PERCEPTION HIERARCHY

Figure 18: High level architecture diagram for the subnetwork of a human-like general intelligence focused on action. From [GIW12]

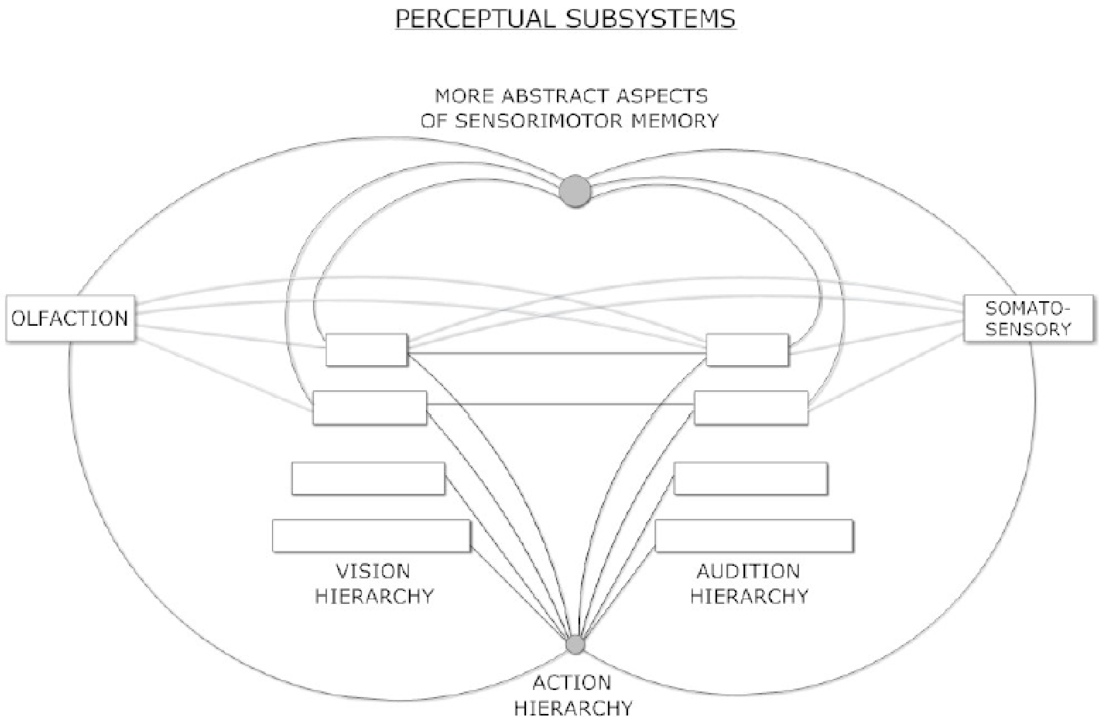


Figure 19: High level architecture diagram for the subnetwork of a human-like general intelligence focused on perception From [GIW12]

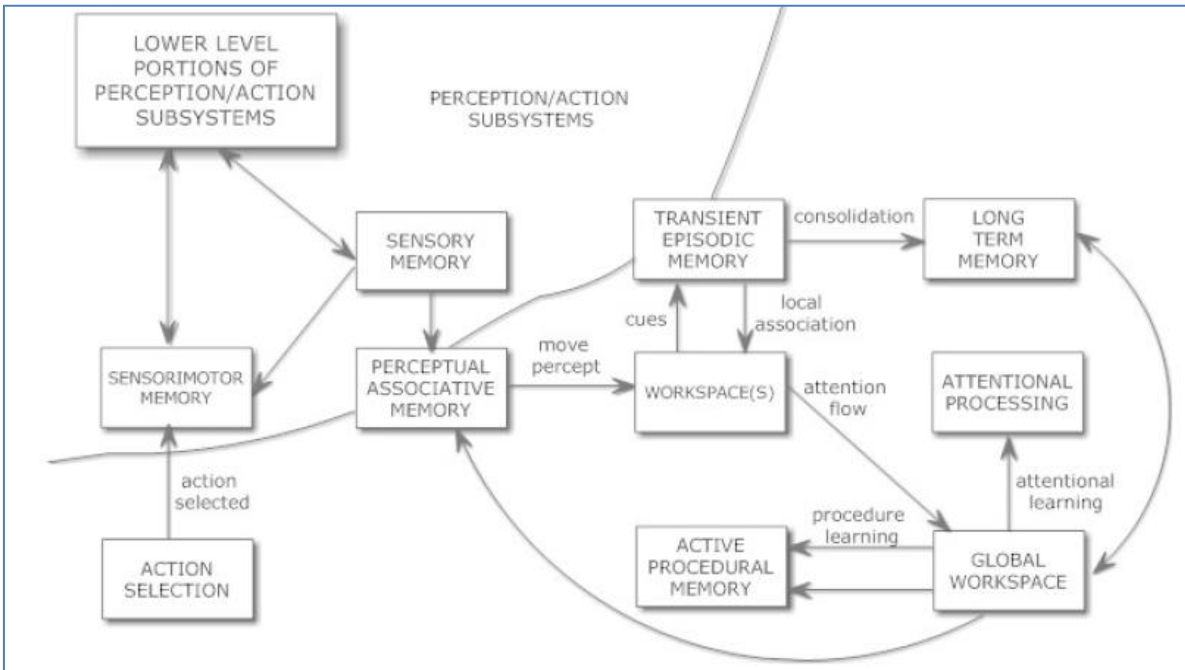


Figure 20: High level architecture diagram for the subnetwork of a human-like general intelligence concerned centrally with working memory. Inspired by the work of Stan Franklin on the LIDA cognitive architecture, among other sources. From [GIW12]

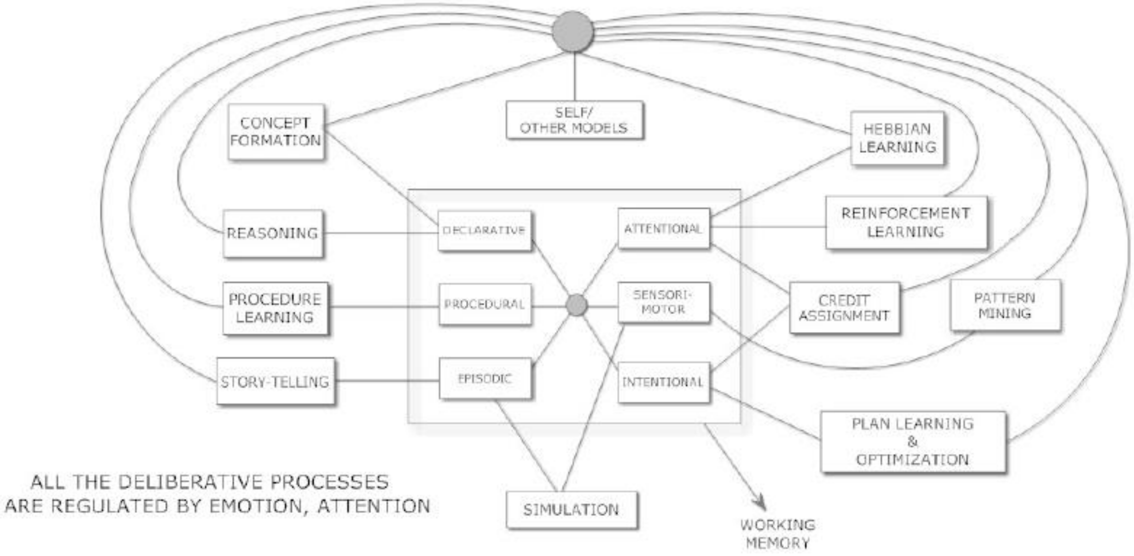


Figure 21: High level architecture diagram for the subnetwork of a human-like general intelligence focused on long-term memory and closely associated reasoning and learning processes. From [GIW12]

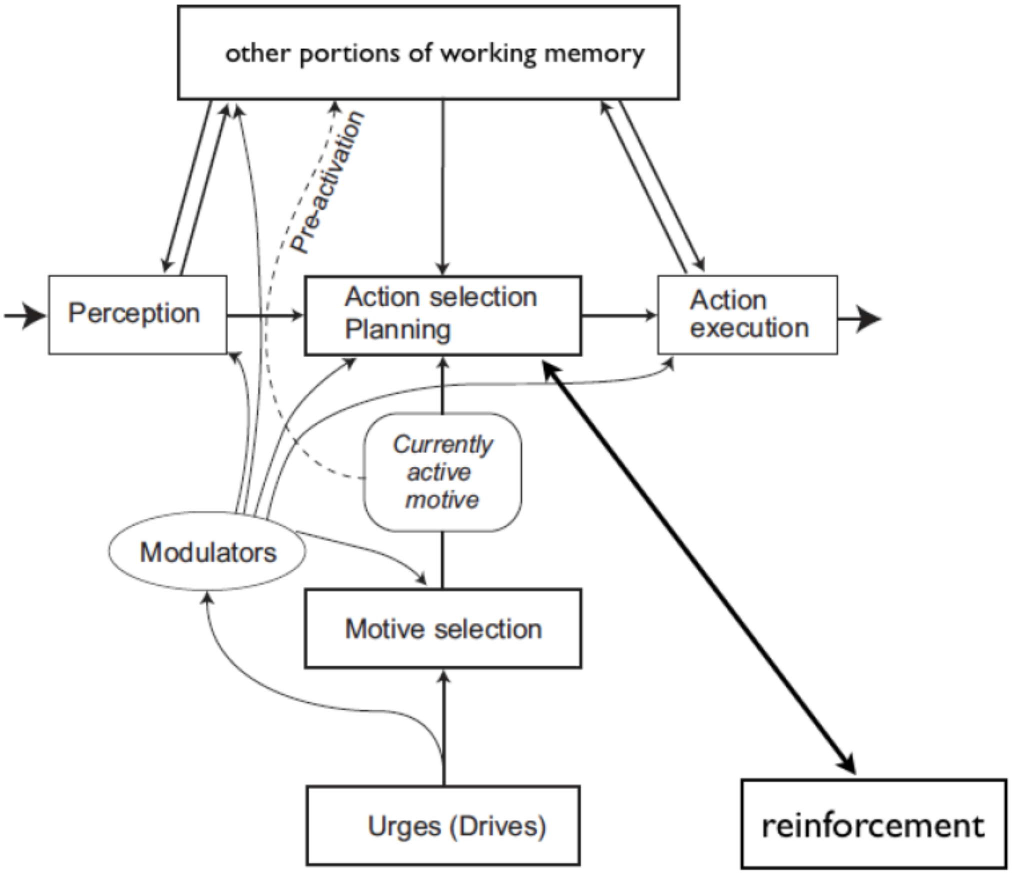


Figure 22: High level architecture diagram for the subnetwork of a human-like general intelligence focused on motivation. Inspired by the work of Joscha Bach on the Psi cognitive model along with other sources. From [GIW12]

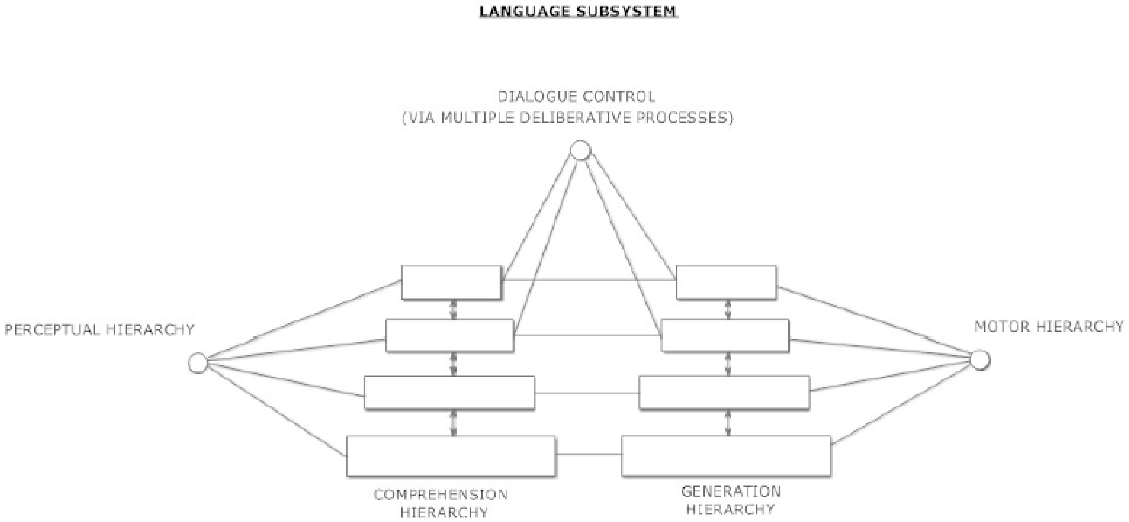


Figure 23: High level architecture diagram for the subnetwork of a human-like general intelligence focused on natural language processing. From [GIW12]

OpenCog Hyperon: A Framework for AGI

Based on **[Goertzel et al. (2023)]** Ben Goertzel, Vitaly Bogdanov, Michael Duncan, Deborah Duong, Zarathustra Goertzel, Jan Horlings, Matthew Ikle', Lucius Greg Meredith, Alexey Potapov, Andre' Luiz de Senna, Hedra Seid Andres Suarez, Adam Vandervorst, Robert Werko. OpenCog Hyperon: A Framework for AGI at the Human Level and Beyond. arXiv:2310.18318v1 [cs.AI] 19 Sep 2023

Contents

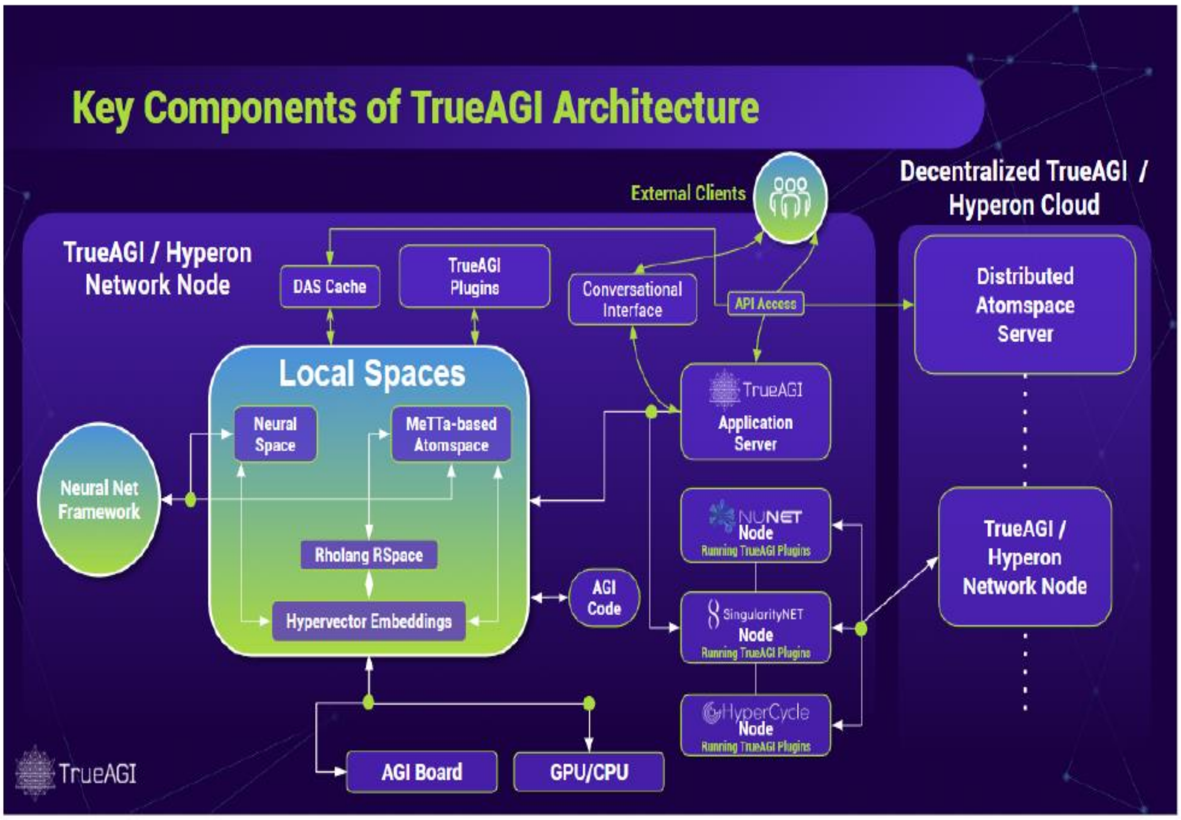

Key Components of TrueAGI Architecture
External Clients
Decentralized TrueAGI / Hyperon Cloud
TrueAGI / Hyperon Network Node
DAS Cache
TrueAGI Plugins
Conversational Interface
API Access
Distributed Atomspace Server
Local Spaces
Neural Space
MeTTa-based Atomspace
Rholang RSpace
Hypervector Embeddings
Neural Net Framework
TrueAGI Application Server
NUNET Node Running TrueAGI Plugins
SingularityNET Node Running TrueAGI Plugins
HyperCycle Node Running TrueAGI Plugins
AGI Code
TrueAGI / Hyperon Network Node
AGI Board
GPU/CPU
TrueAGI

Figure 1: High-level illustration of key components in Hyperon architecture, including integration into TrueAGI application framework.

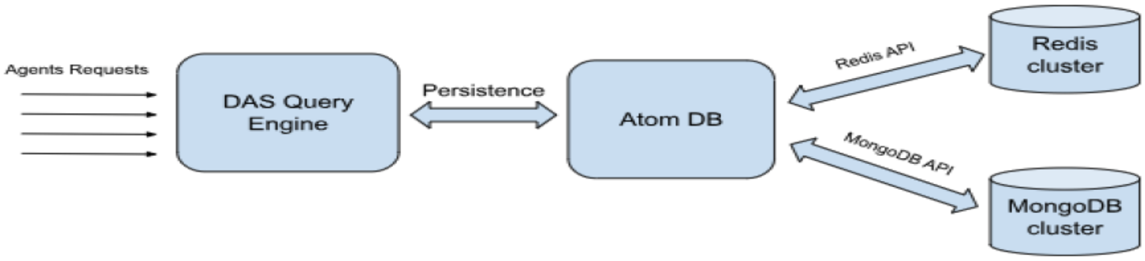


Figure 2: High level architecture of Hyperon Distributed Atomspace

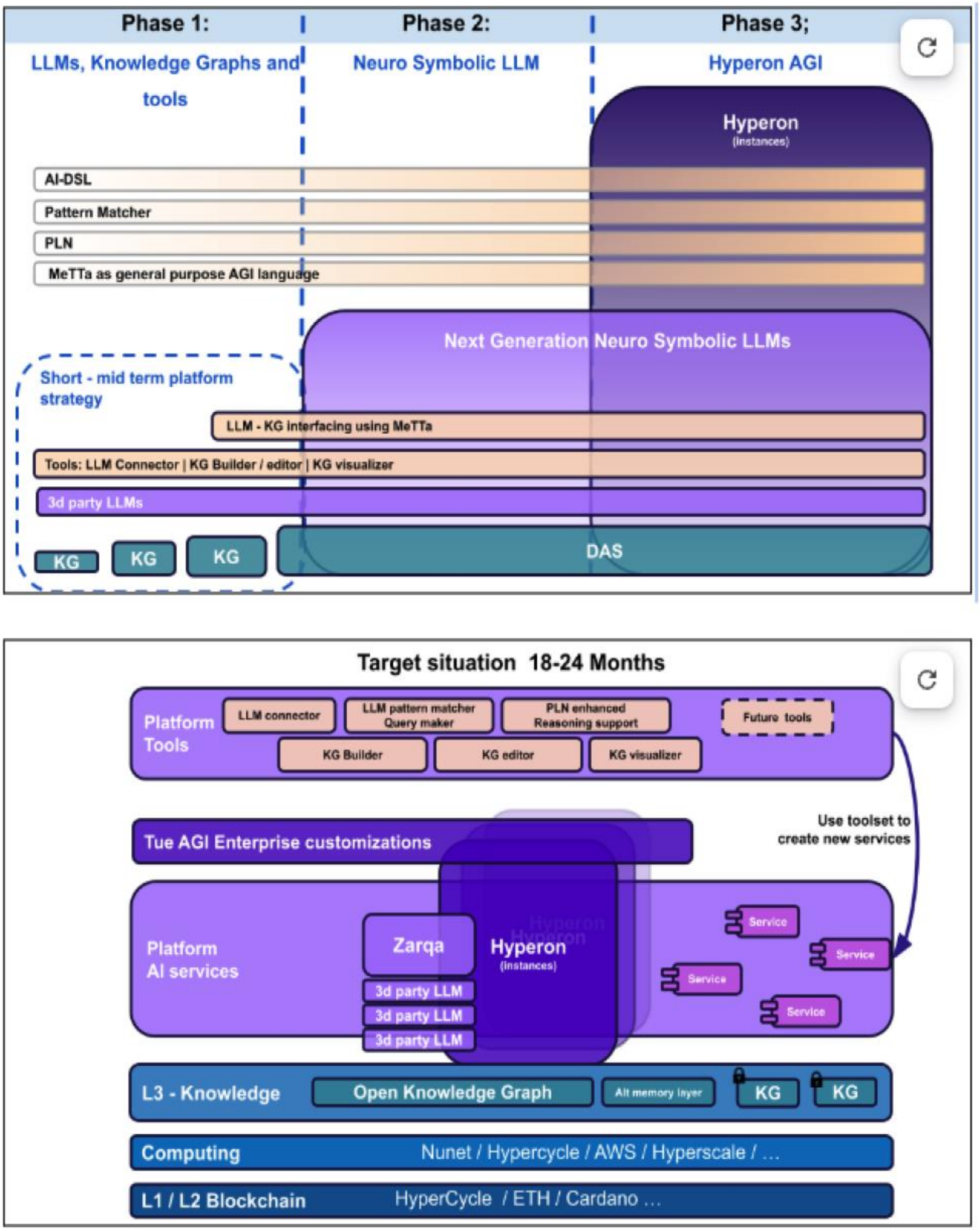

Phase 1:
LLMs, Knowledge Graphs and tools
Phase 2:
Neuro Symbolic LLM
Phase 3;
Hyperon AGI
Hyperon
(instances)
AI-DSL
Pattern Matcher
PLN
MeTTa as general purpose AGI language
Next Generation Neuro Symbolic LLMs
Short - mid term platform strategy
LLM - KG interfacing using MeTTa
Tools: LLM Connector | KG Builder / editor | KG visualizer
3d party LLMs
KG
KG
KG
DAS
Target situation 18-24 Months
Platform Tools
LLM connector
LLM pattern matcher Query maker
PLN enhanced Reasoning support
Future tools
KG Builder
KG editor
KG visualizer
Use toolset to create new services
Tue AGI Enterprise customizations
Platform AI services
Zarqa
3d party LLM
3d party LLM
3d party LLM
Hyperon
(instances)
Service
Service
Service
Service
L3 - Knowledge
Open Knowledge Graph
Alt memory layer
KG
KG
Computing
Nunet / Hypercycle / AWS / Hyperscale / ...
L1 / L2 Blockchain
HyperCycle / ETH / Cardano ...

Figure 3: Rough tentative roadmap for some of the key development initiatives regarding the SingularityNET platform

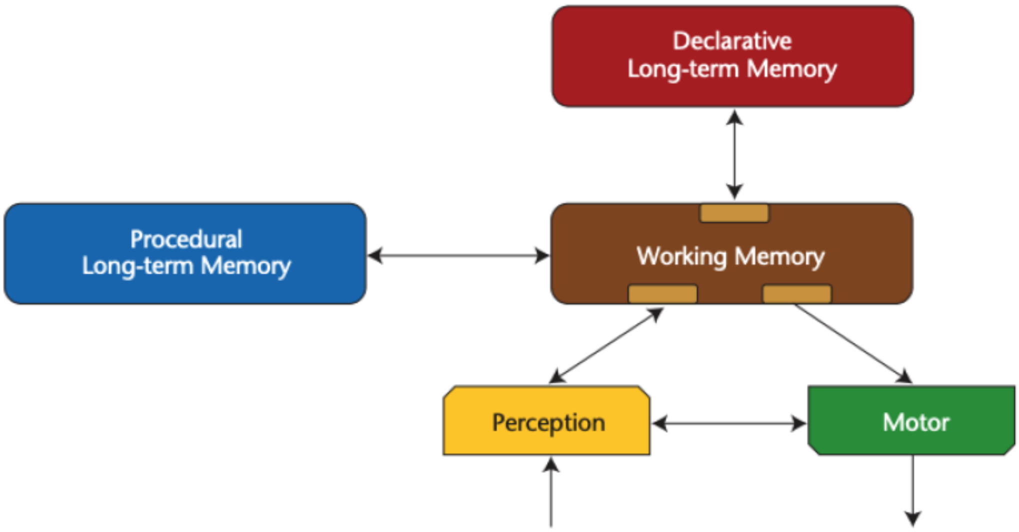


Figure 16: Standard Model of Mind: High-Level Cognitive Architecture

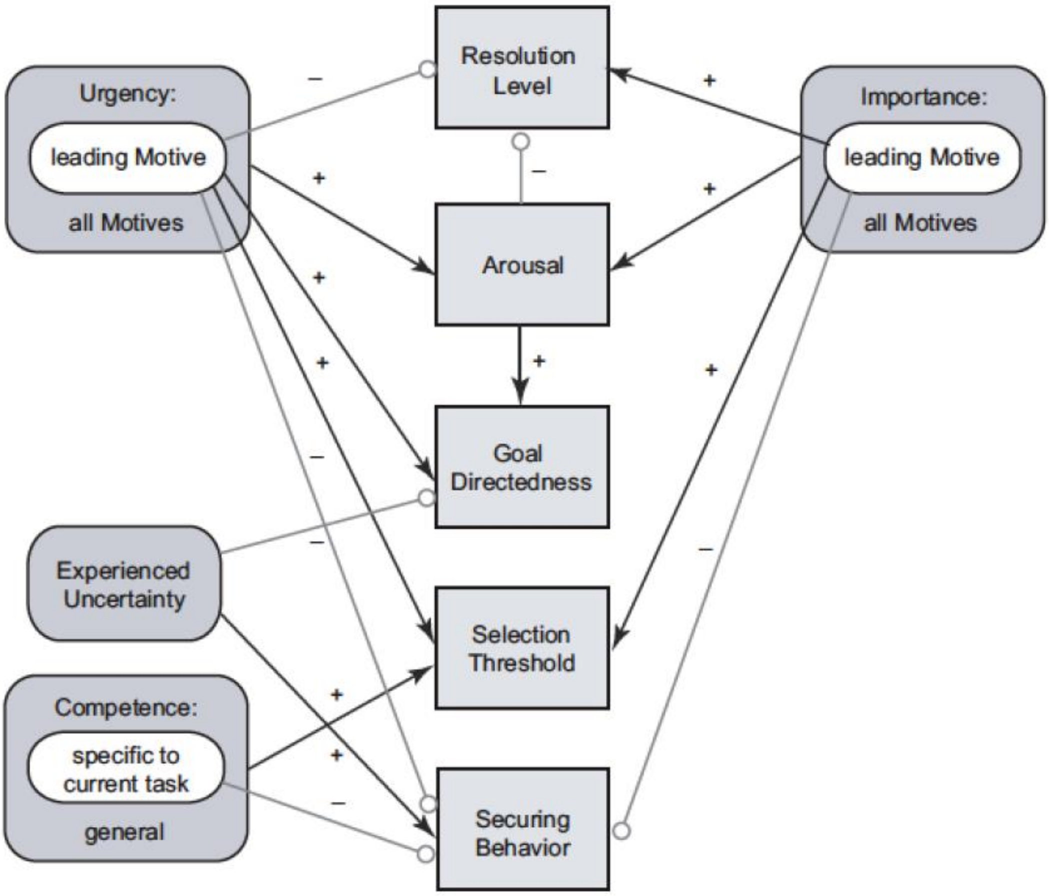


Figure 18: Graphical illustration of the basic dynamics of human-like emotion according to Joscha Bach's Psi model

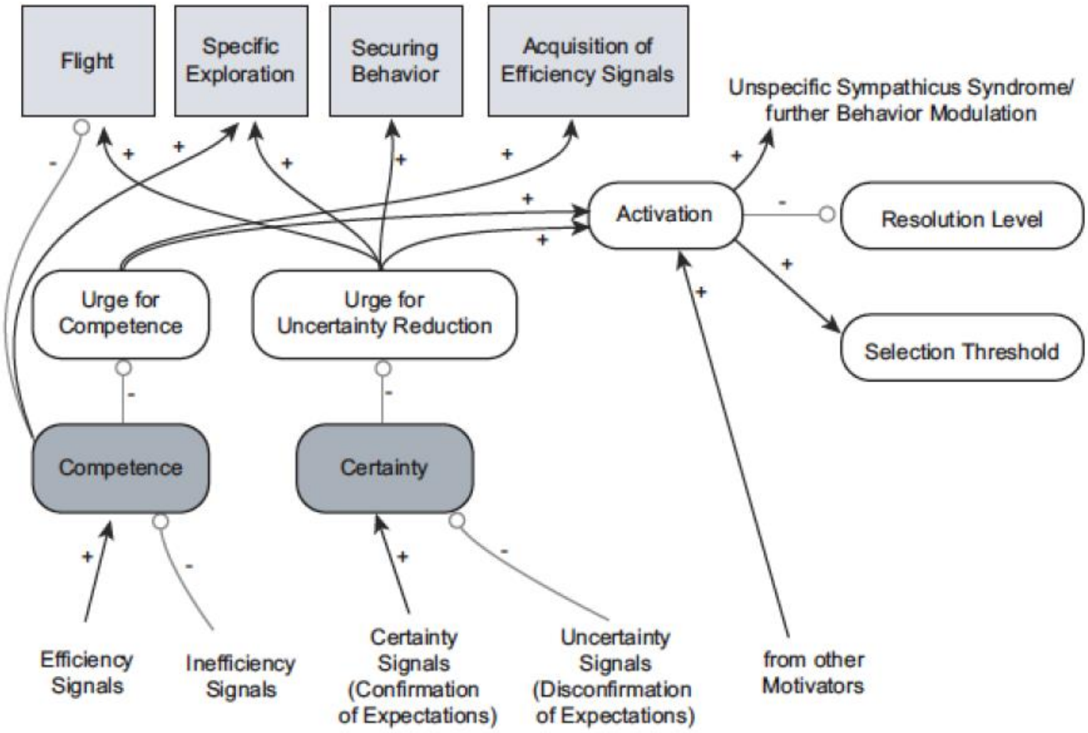

Flight
Specific Exploration
Securing Behavior
Acquisition of Efficiency Signals
Unspecific Sympathicus Syndrome/ further Behavior Modulation
Activation
Resolution Level
Selection Threshold
Urge for Competence
Urge for Uncertainty Reduction
Competence
Certainty
Efficiency Signals
Inefficiency Signals
Certainty Signals (Confirmation of Expectations)
Uncertainty Signals (Disconfirmation of Expectations)
from other Motivators
+
-

Figure 19: Graphical illustration of the process via which various parameters ("modulators") guide human-like action and emotion according to Joscha Bach's Psi model

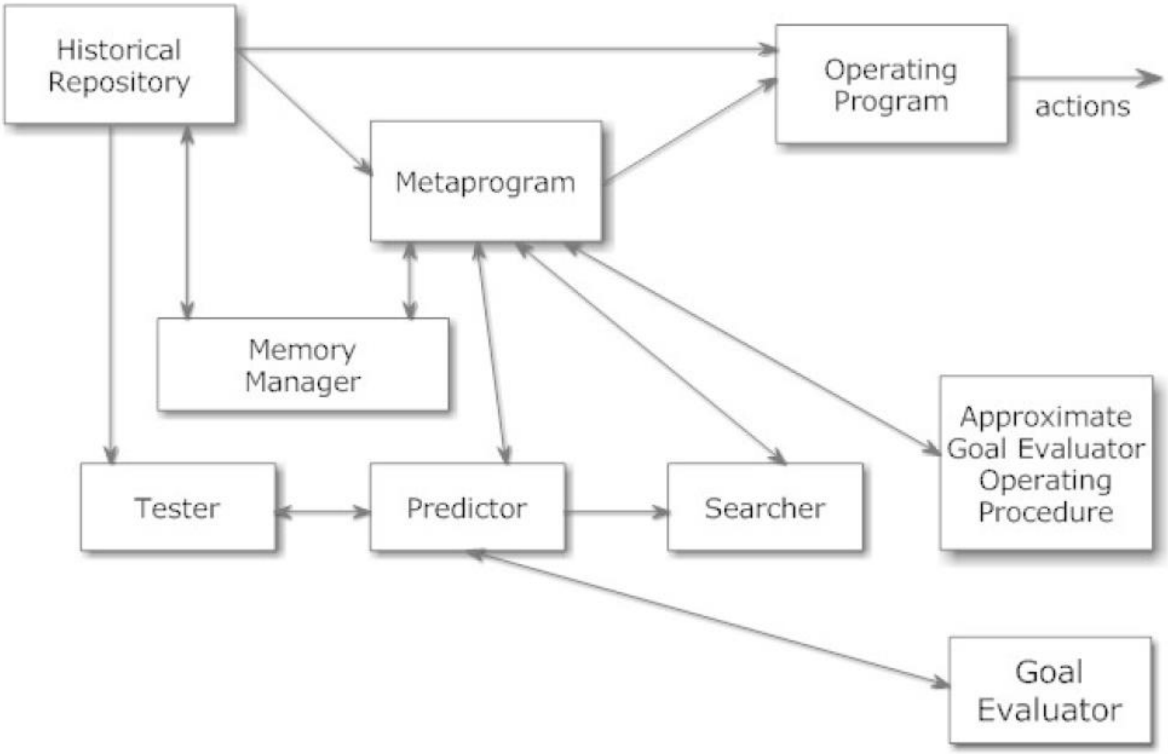


Figure 20: Sketch of GOLEM meta-architecture for relatively safe/reliable self-modifying AI [Goe14b]

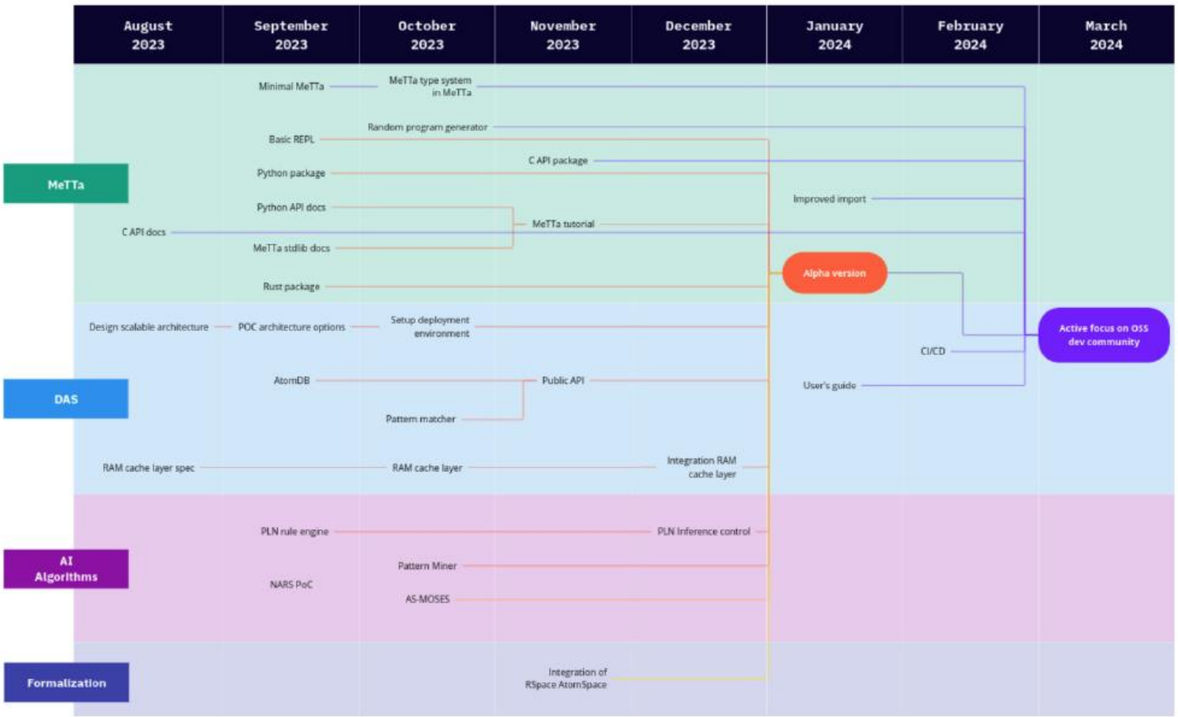
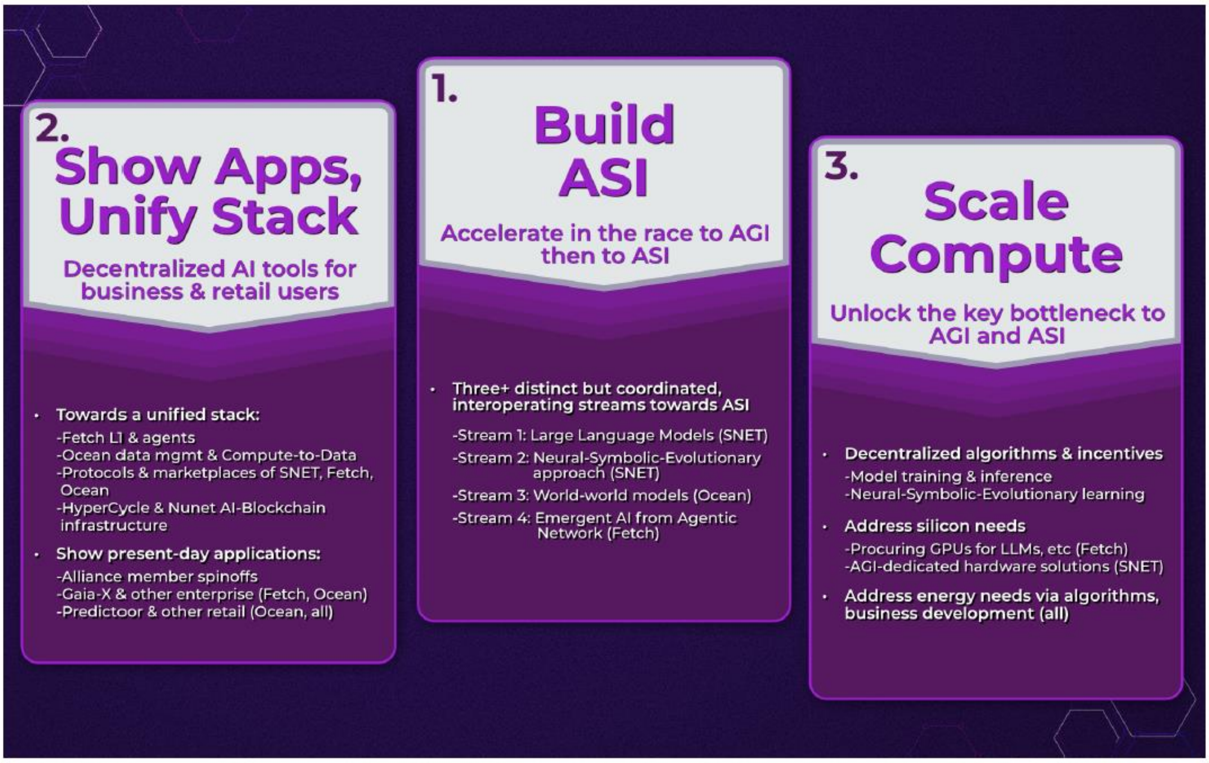


Figure 22: Rough tentative roadmap for some of the key development initiatives regarding core OpenCog Hyperon technology (not attempting to cover specific applications, demos or commercialization efforts)

Decentralized Artificial Superintelligence

Based on **[ASI Alliance (2024)]** Artificial Superintelligence (ASI) Alliance Vision Paper**.** Building Decentralized Artificial Superintelligence**.** SingularityNET, Fetch.ai, Ocean Protocol - Apr 2024



2. Show Apps, Unify Stack

Decentralized AI tools for business & retail users

- Towards a unified stack:
  -Fetch L1 & agents
  -Ocean data mgmt & Compute-to-Data
  -Protocols & marketplaces of SNET, Fetch, Ocean
  -HyperCycle & Nunet AI-Blockchain infrastructure
- Show present-day applications:
  -Alliance member spinoffs
  -Gaia-X & other enterprise (Fetch, Ocean)
  -Predictoor & other retail (Ocean, all)

1. Build ASI

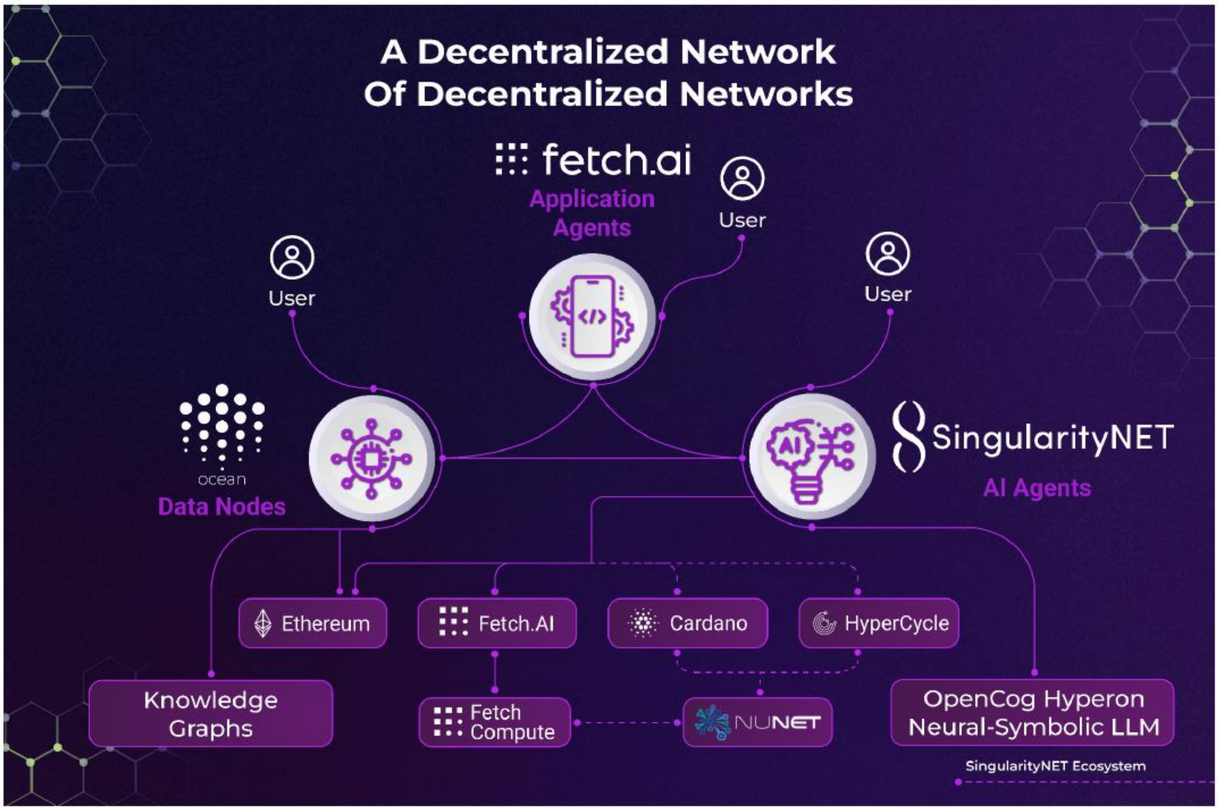
Accelerate in the race to AGI then to ASI

- Three+ distinct but coordinated, interoperating streams towards ASI
  -Stream 1: Large Language Models (SNET)
  -Stream 2: Neural-Symbolic-Evolutionary approach (SNET)
  -Stream 3: World-world models (Ocean)
  -Stream 4: Emergent AI from Agentic Network (Fetch)

3. Scale Compute

Unlock the key bottleneck to AGI and ASI

- Decentralized algorithms & incentives
  -Model training & inference
  -Neural-Symbolic-Evolutionary learning
- Address silicon needs
  -Procuring GPUs for LLMs, etc (Fetch)
  -AGI-dedicated hardware solutions (SNET)
- Address energy needs via algorithms, business development (all)



Key findings for our Project

- Theoretical model, based on several different cognitological models/approaches
- Cybernetic control algorithms and systems
- Semiotic approaches
- Internal space and modelling
- Framework for Multiagent MAS AGI/ASI system
- Realistic and promising ASI (not only AGI) Project
- ASI Roadmap

T. The Thousand Brains Project

We will devote this chapter to significant and promising AGI Project, presented in the short Overview - **[Clay, Leadholm & Hawkins (2024)]** Viviane Clay, Niels Leadholm, and Jeff Hawkins. The Thousand Brains Project. Numenta, 2024

Contents

Key References

- [Hawkins, Jeff & Ahmad (2016)] - "**Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex"**
- [Hawkins, Jeff, Ahmad & Cui (2017)] - **"A Theory of How Columns in the Neocortex Enable Learning the Structure of the World"**
- [Hawkins, Jeff et al. (2019)] - **"A framework for intelligence and cortical function based on grid cells in the neocortex"**
- [Ahmad, Subutai & Scheinkman (2019)] - **"How Can We Be So Dense? The Robustness of Highly Sparse Representations"**
- [Lewis, Marcus et al. (2019)] - **"Locations in the neocortex: A theory of sensorimotor object recognition using cortical grid cells"**

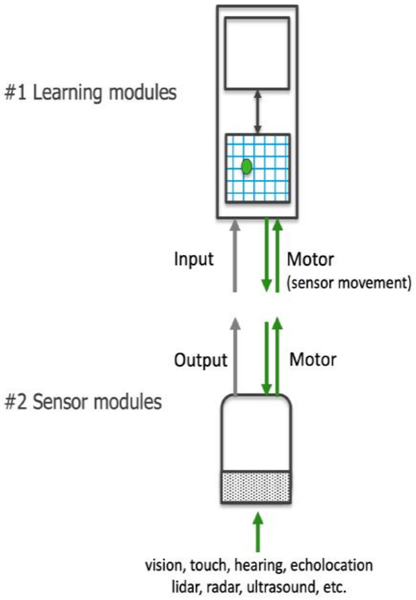


Figure 1.1: Sensor modules receive and process the raw sensory input. This is then communicated via a common communication protocol to a learning module which uses this to learn and recognize models of anything in the environment.

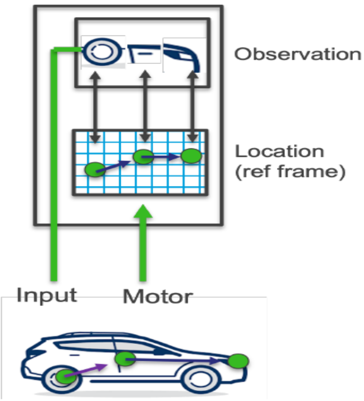


Figure 1.2: Learning modules learn structured models through sensorimotor interaction, using reference frames. They model how incoming features are arranged relative to each other in space and time.

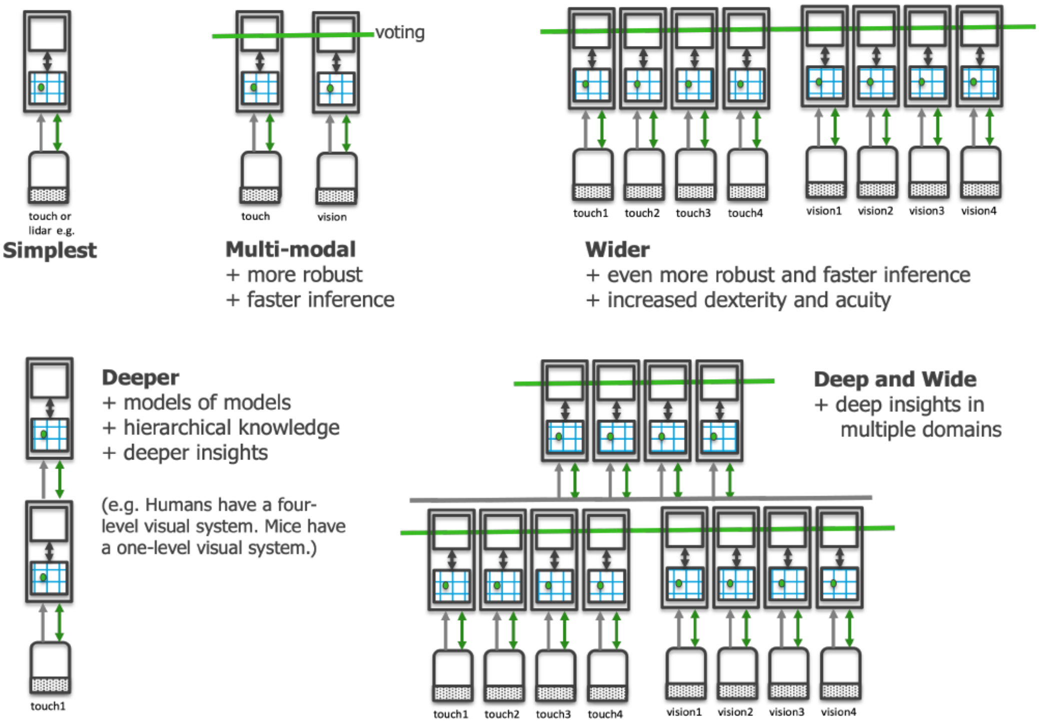


Figure 1.4: By using a common communication protocol between sensor modules and learning modules, the system can easily be scaled in multiple dimensions. This provides a straightforward way for dealing with multiple sensory inputs from multiple modalities. Using multiple learning modules next to each other can improve robustness through votes between them. Additionally, stacking learning modules on top of each other allows for more complex, hierarchical processing of inputs and modeling compositional objects.

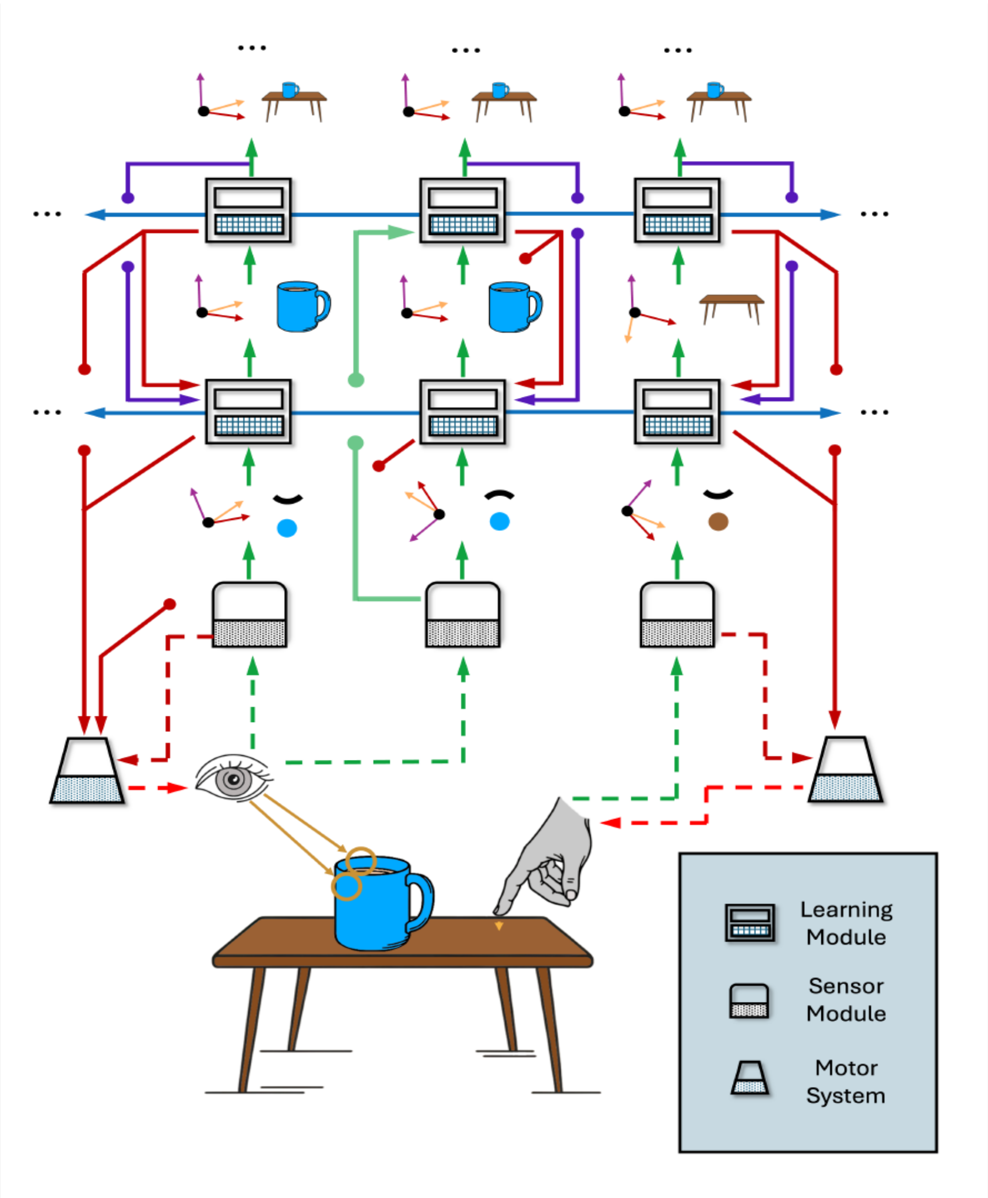


Figure 1.6: High-level overview the Architecture with all the main conceptual components mirroring figure 1.4 applied to a concrete example. Green lines indicate the main flow of information up the hierarchy. Purple lines show top-down connections, biasing the lower-level learning modules. Light blue lines show lateral voting connections. Red lines show the communication of goal states which eventually translate into motor commands in the motor system. Every LM has a direct motor output. Information communicated along solid lines follows the CCP (contains features and pose). Discontinuations in the diagram are marked with dots on line-ends. Dashed lines are the interface of the system with the world and subcortical compute units and do not need to follow the CCP. Green dashed lines communicate raw sensory input from sensors. Red dashed lines communicate motor commands to the actuators. The dark red dashed lines send sensory information directly to the motor system and implement a fast reflex loop for purely input-driven policies. The large, semi-transparent green arrow is an example of a connection carrying sensory outputs from a larger receptive field directly to the higher-level LM.

Key findings for our Project

- **Theoretical cognitological model**
- **Cybernetic control algorithms**
- **Embodied, sensorimotor system**
- **Internal space and modelling**
- **Platform (Framework) for multiagent MAS and modular AGI system**
- **Realistic and promising AGI Project**
- **AGI Roadmap**

U. Generative AI for Self-Adaptive Systems

We will devote this chapter to interesting and promising paper, presented in the short Overview **[Li et al. (2024a)]** Jialong Li, Mingyue Zhang, Nianyu Li, Danny Weyns, Zhi Jin, and Kenji Tei. 2024. Generative AI for Self-Adaptive Systems: State of the Art and Research Roadmap. Article in ACM Transactions on Autonomous and Adaptive Systems · September 2024

Self-adaptive systems (SASs) are designed to handle changes and uncertainties through a feedback loop with four core functionalities: monitoring, analyzing, planning, and execution. Recently, generative artificial intelligence (GenAI), especially the area of large language models, has shown impressive performance in data comprehension and logical reasoning. These capabilities are highly aligned with the functionalities required in SASs, suggesting a strong potential to employ GenAI to enhance SASs.

This study outlines a research roadmap that highlights the challenges of integrating GenAI into SASs. The roadmap starts with outlining key research challenges that need to be tackled to exploit the potential for applying GenAI in the field of SAS. The roadmap concludes with a practical reflection, elaborating on current shortcomings of GenAI and proposing possible mitigation strategies.

Five key references for this Paper (reference list total 20+ pages!):

- [Weyns (2020)] - **An Introduction to Self-adaptive Systems: A Contemporary Software Engineering Perspective**
- [Weyns et al. (2022)] - **The vision of self-evolving computing systems**
- [Kephart & Chess (2003)] - **The vision of autonomic computing**
- [Andersson et al. (2009)] - **Modeling Dimensions of Self-Adaptive Software Systems**
- [Ferreira, Silva & Martins (2024)] - **Organizing a Society of Language Models: Structures and Mechanisms for Enhanced Collective Intelligence**

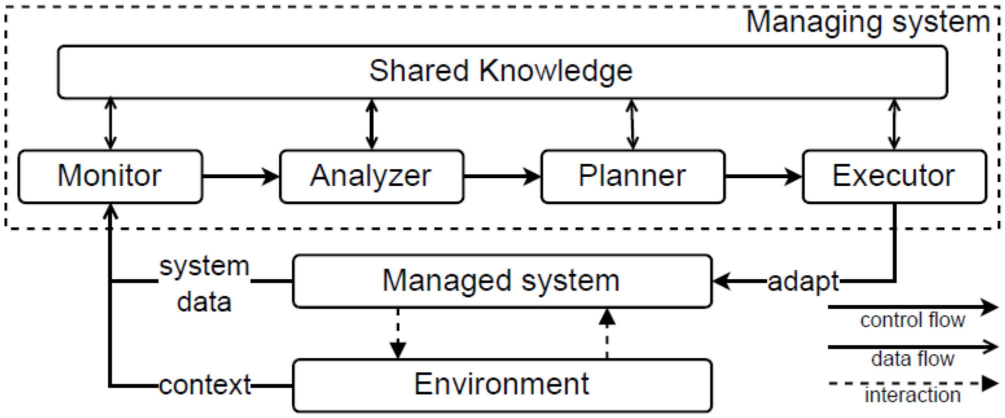


Fig. 2. Self-adaptive System with MAPE-K Feedback Loop [Andersson et al. 2009].

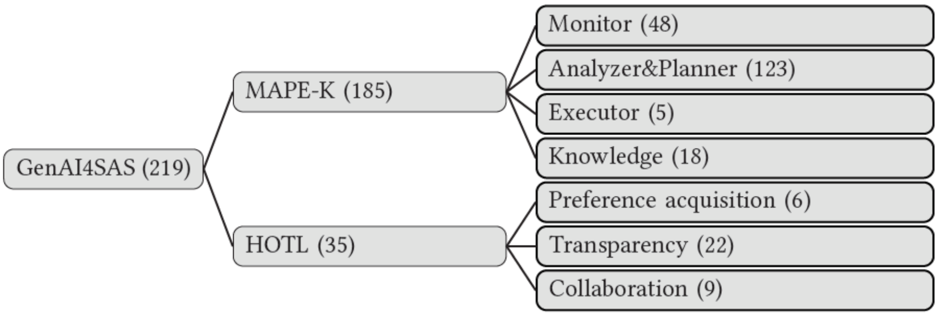


Fig. 3. Literature Categorization Overview. One piece of literature may be involved in multiple categories.

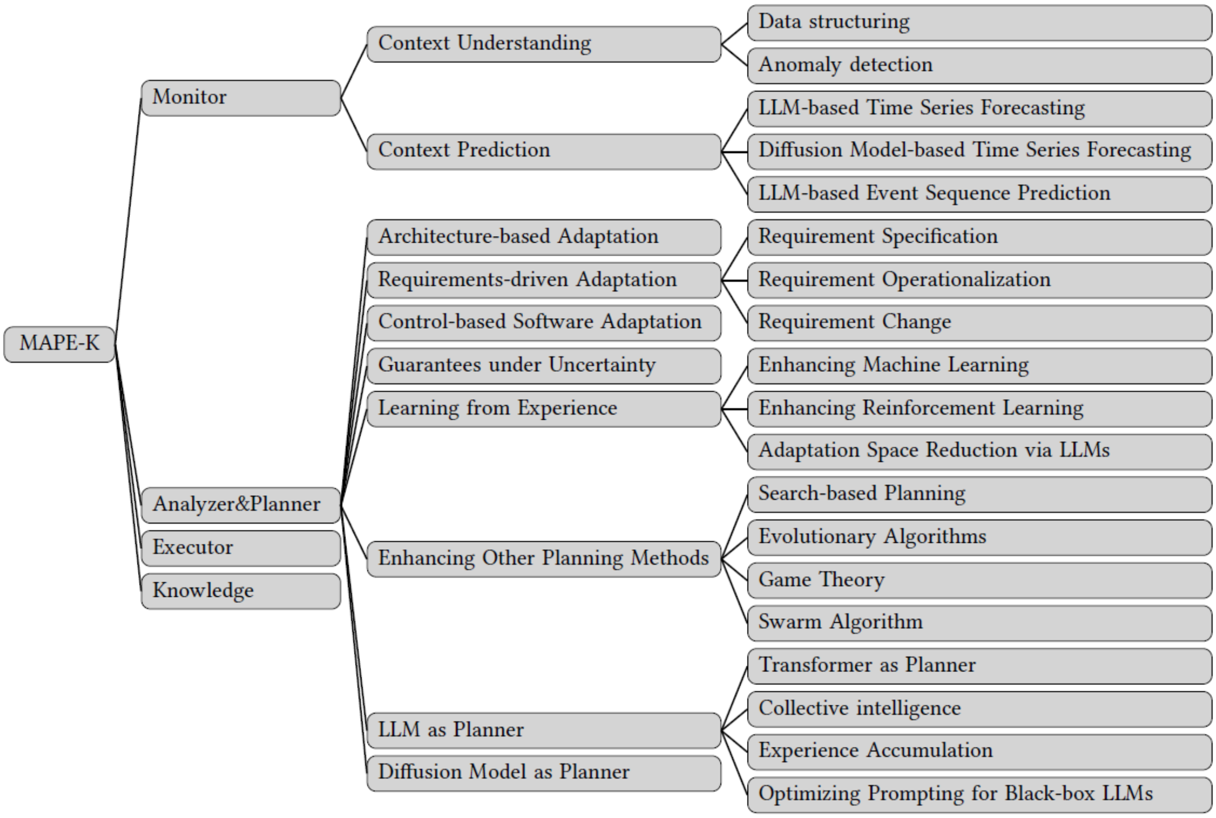

MAPE-K
Monitor
Analyzer&Planner
Executor
Knowledge
Context Understanding
Context Prediction
Architecture-based Adaptation
Requirements-driven Adaptation
Control-based Software Adaptation
Guarantees under Uncertainty
Learning from Experience
Enhancing Other Planning Methods
LLM as Planner
Diffusion Model as Planner
Data structuring
Anomaly detection
LLM-based Time Series Forecasting
Diffusion Model-based Time Series Forecasting
LLM-based Event Sequence Prediction
Requirement Specification
Requirement Operationalization
Requirement Change
Enhancing Machine Learning
Enhancing Reinforcement Learning
Adaptation Space Reduction via LLMs
Search-based Planning
Evolutionary Algorithms
Game Theory
Swarm Algorithm
Transformer as Planner
Collective intelligence
Experience Accumulation
Optimizing Prompting for Black-box LLMs

Fig. 4. Overview of Empowerment of MAPE-K Modules via GenAI.

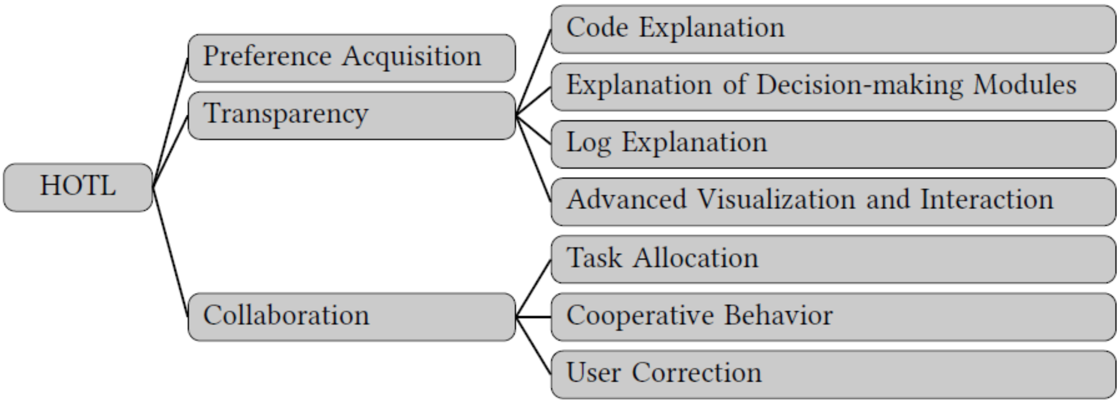


Fig. 5. Overview of Empowerment of Human-on-the-loop via GenAI.

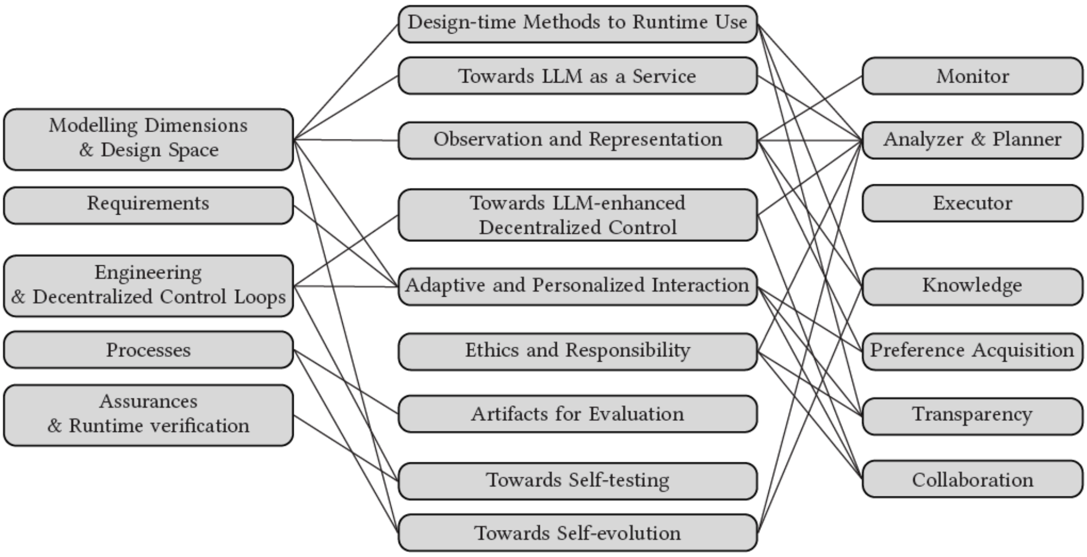

Modelling Dimensions & Design Space
Requirements
Engineering & Decentralized Control Loops
Processes
Assurances & Runtime verification
Design-time Methods to Runtime Use
Towards LLM as a Service
Observation and Representation
Towards LLM-enhanced Decentralized Control
Adaptive and Personalized Interaction
Ethics and Responsibility
Artifacts for Evaluation
Towards Self-testing
Towards Self-evolution
Monitor
Analyzer & Planner
Executor
Knowledge
Preference Acquisition
Transparency
Collaboration

Fig. 6. Left-hand side: key software engineering aspects that need to be considered in the design and realization of self-adaptive systems. Middle: challenges of employing GenAI and LLMs in particular in self-adaptive systems. Right-hand side: primary functions that are involved in self-adaptation with an emphasis on MAPE-K and HOTL. Mapping expresses the relationships between the concepts and challenges.

Key findings for our Project

- **Self-adaptive SAS and self-evolving Systems - SO**
- **Cybernetic control algorithms**
- **Control and meta-control functions**
- **MASs and collective Intelligence**
- **R&D Roadmap**

V. A Landscape of Consciousness

Interesting and wide overview (*Reference list total 25+ pages!)* paper, presented in the short Overview **[Kuhn (2024)]** R. L. Kuhn, “A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications,” Progress in Biophysics and Molecular Biology, 190, 28–169 [Aug.] (2024)

1. MATERIALISM THEORIES

1.1. Philosophical Theories

1.1.1.Eliminative Materialism / Illusionism
1.1.2.Epiphenomenalism
1.1.3.Functionalism
1.1.4.Emergence
1.1.5.Mind-Brain Identity Theory
1.1.6.Searle’s Biological Naturalism
1.1.7.Block’s Biological Reductionism
1.1.8.Flanagan’s Constructive Naturalism
1.1.9.Papineau’s Mind-Brain Identity
1.1.10. Goldstein’s Mind-Body Problem
1.1.11. Hardcastle’s Argument Against Materialism Skeptics
1.1.12. Stoljar’s Epistemic View and Non-Standard Physicalism

1.2. Neurobiological Theories

1.2.1.Edelman’s Neural Darwinism and Reentrant Neural Circuitry
1.2.2.Crick and Koch’s Neural Correlates of Consciousness
1.2.3.Baars’s and Dehaene’s Global Workspace Theory (*GWT*)
1.2.4.Dennett’s Multiple Drafts Model
1.2.5.Minsky’s Society of Mind
1.2.6.Graziano’s Attention Schema Theory
1.2.7.Prinz’s Neurofunctionalism: Attention Engenders Experience
1.2.8.Sapolsky’s Hard Incompatibilism
1.2.9.Mitchell’s Free Agents
1.2.10. Bach’s Cortical Conductor Theory
1.2.11. Brain Circuits and Cycles Theories
1.2.12. Northoff’s Temporo-Spatial Sentience
1.2.13. Bunge’s Emergent Materialism
1.2.14. Hirstein’s Mindmelding

1.3. Electromagnetic (*EM*) Field Theories

1.3.1.Jones’s Electromagnetic Fields
1.3.2.Pockett’s Conscious and Non-Conscious Patterns
1.3.3.McFadden’s Conscious Electromagnetic Information Theory
1.3.4.Ephaptic Coupling
1.3.5.Ambron’s Local Field Potentials and Electromagnetic Waves
1.3.6.Llinas’s Mindness State of Oscillations
1.3.7.Zhang’s Long-Distance Light-Speed Telecommunications

1.4. Computational and Informational Theories

1.4.1.Computational Theories
1.4.2.Grossberg’s Adaptive Resonance Theory
1.4.3.Complex Adaptive Systems Models

10.4. Eagleman’s Possibilianismˆ

10.5. Tallis’s Anti-Neuromania Skepticism

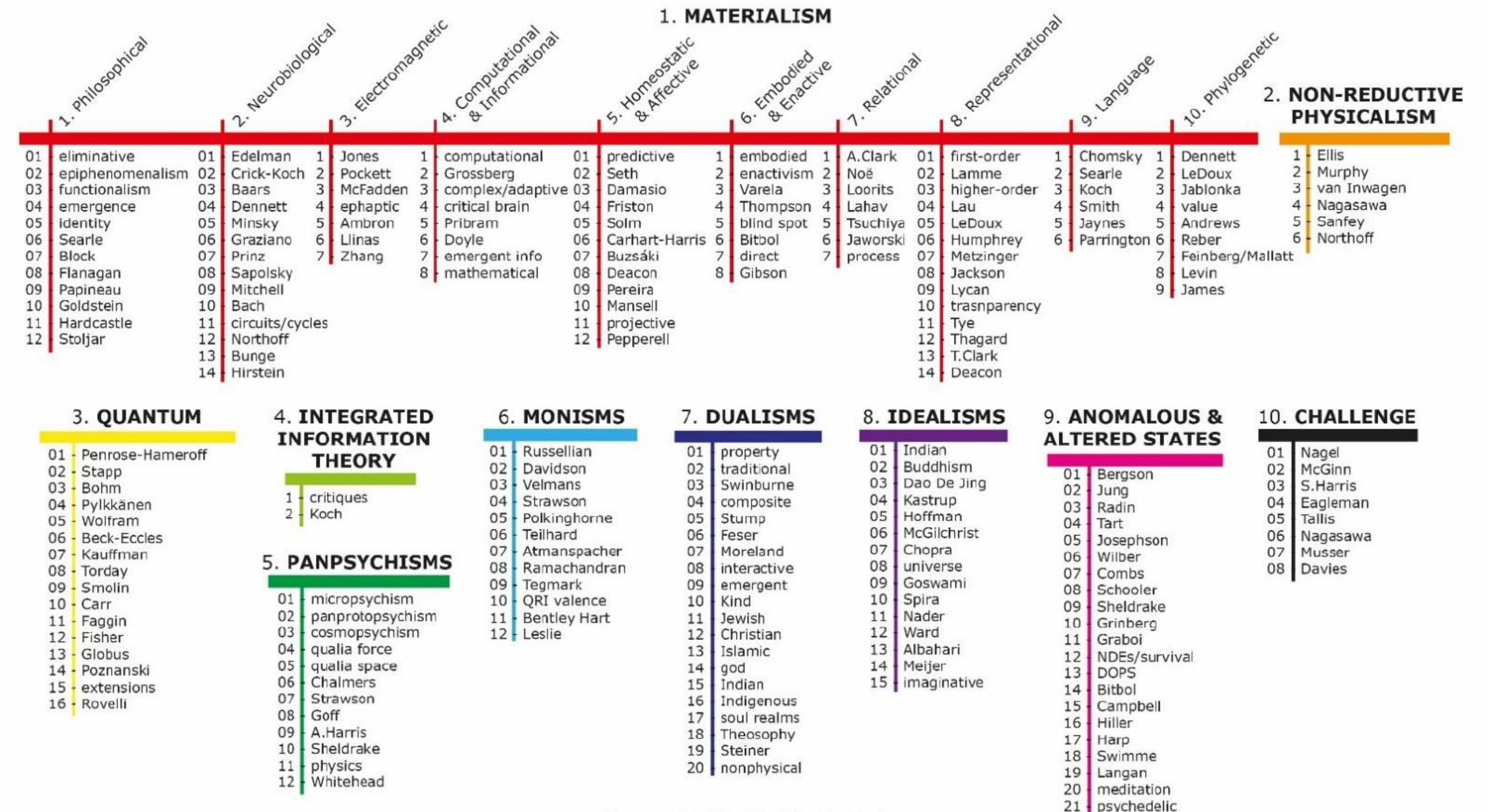
10.6. Nagasawa’s Mind-Body in an Infinitely Decomposable Universe

10.7. Musser’s “Is It Really So Hard?”

10.8. Davies’s Consciousness in the Cosmos

Key findings for our Project

Wide and comprehensive overview of Theories and Models of Consciousness with well-structured and systematic Taxonomy.



1. MATERIALISM

1. Philosophical
01 eliminative
02 epiphenomenalism
03 functionalism
04 emergence
05 identity
06 Searle
07 Block
08 Flanagan
09 Papineau
10 Goldstein
11 Hardcastle
12 Stoljar

2. Neurobiological
01 Edelman
02 Crick-Koch
03 Baars
04 Dennett
05 Minsky
06 Graziano
07 Prinz
08 Sapolsky
09 Mitchell
10 Bach
11 circuits/cycles
12 Northoff
13 Bunge
14 Hirstein

3. Electromagnetic
1 Jones
2 Pockett
3 McFadden
4 ephaptic
5 Ambron
6 Llinas
7 Zhang

4. Computational & Informational
1 computational
2 Grossberg
3 complex/adaptive
4 critical brain
5 Pribram
6 Doyle
7 emergent info
8 mathematical

5. Homeostatic & Affective
01 predictive
02 Seth
03 Damasio
04 Friston
05 Solm
06 Carhart-Harris
07 Buzsáki
08 Deacon
09 Pereira
10 Mansell
11 projective
12 Pepperell

6. Embodied & Enactive
1 embodied
2 enactivism
3 Varela
4 Thompson
5 blind spot
6 Bitbol
7 direct
8 Gibson

7. Relational
1 A.Clark
2 Noë
3 Loorits
4 Lahav
5 Tsuchiya
6 Jaworski
7 process

8. Representational
01 first-order
02 Lamme
03 higher-order
04 Lau
05 LeDoux
06 Humphrey
07 Metzinger
08 Jackson
09 Lycan
10 trasnparency
11 Tye
12 Thagard
13 T.Clark
14 Deacon

9. Language
1 Chomsky
2 Searle
3 Koch
4 Smith
5 Jaynes
6 Parrington

10. Phylogenetic
1 Dennett
2 LeDoux
3 Jablonka
4 value
5 Andrews
6 Reber
7 Feinberg/Mallatt
8 Levin
9 James

2. NON-REDUCTIVE PHYSICALISM
1 Ellis
2 Murphy
3 van Inwagen
4 Nagasawa
5 Sanfey
6 Northoff

3. QUANTUM
01 Penrose-Hameroff
02 Stapp
03 Bohm
04 Pylkkänen
05 Wolfram
06 Beck-Eccles
07 Kauffman
08 Torday
09 Smolin
10 Carr
11 Faggin
12 Fisher
13 Globus
14 Poznanski
15 extensions
16 Rovelli

4. INTEGRATED INFORMATION THEORY
1 critiques
2 Koch

5. PANPSYCHISMS
01 micropsychism
02 panprotopsychism
03 cosmopsychism
04 qualia force
05 qualia space
06 Chalmers
07 Strawson
08 Goff
09 A.Harris
10 Sheldrake
11 physics
12 Whitehead

6. MONISMS
01 Russellian
02 Davidson
03 Velmans
04 Strawson
05 Polkinghorne
06 Teilhard
07 Atmanspacher
08 Ramachandran
09 Tegmark
10 QRI valence
11 Bentley Hart
12 Leslie

7. DUALISMS
01 property
02 traditional
03 Swinburne
04 composite
05 Stump
06 Feser
07 Moreland
08 interactive
09 emergent
10 Kind
11 Jewish
12 Christian
13 Islamic
14 god
15 Indian
16 Indigenous
17 soul realms
18 Theosophy
19 Steiner
20 nonphysical

8. IDEALISMS
01 Indian
02 Buddhism
03 Dao De Jing
04 Kastrup
05 Hoffman
06 McGilchrist
07 Chopra
08 universe
09 Goswami
10 Spira
11 Nader
12 Ward
13 Albahari
14 Meijer
15 imaginative

9. ANOMALOUS & ALTERED STATES
01 Bergson
02 Jung
03 Radin
04 Tart
05 Josephson
06 Wilber
07 Combs
08 Schooler
09 Sheldrake
10 Grinberg
11 Graboi
12 NDEs/survival
13 DOPS
14 Bitbol
15 Campbell
16 Hiller
17 Harp
18 Swimme
19 Langan
20 meditation
21 psychedelic

10. CHALLENGE
01 Nagel
02 McGinn
03 S.Harris
04 Eagleman
05 Tallis
06 Nagasawa
07 Musser
08 Davies

Figure created by Alex Gomez-Marin

W. Why Is Anything Conscious?

We will devote this chapter to interesting and significant paper, presented in the short Overview **[Bennett, Welsh & Ciaunica (2024)]** Michael Timothy Bennett, Sean Welsh, Anna Ciaunica. Why Is Anything Conscious? arXiv:2409.14545v2 [cs.AI] 2 Nov 2024

The Paper Authors tackle the hard problem of consciousness taking the naturally selected, selforganising, embodied organism as our starting point. Also, provide a mathematical formalism describing how biological systems self-organise to hierarchically interpret unlabelled sensory information according to valence and specific needs and formally describe the multilayered architecture of self-organisation from rocks to Einstein.

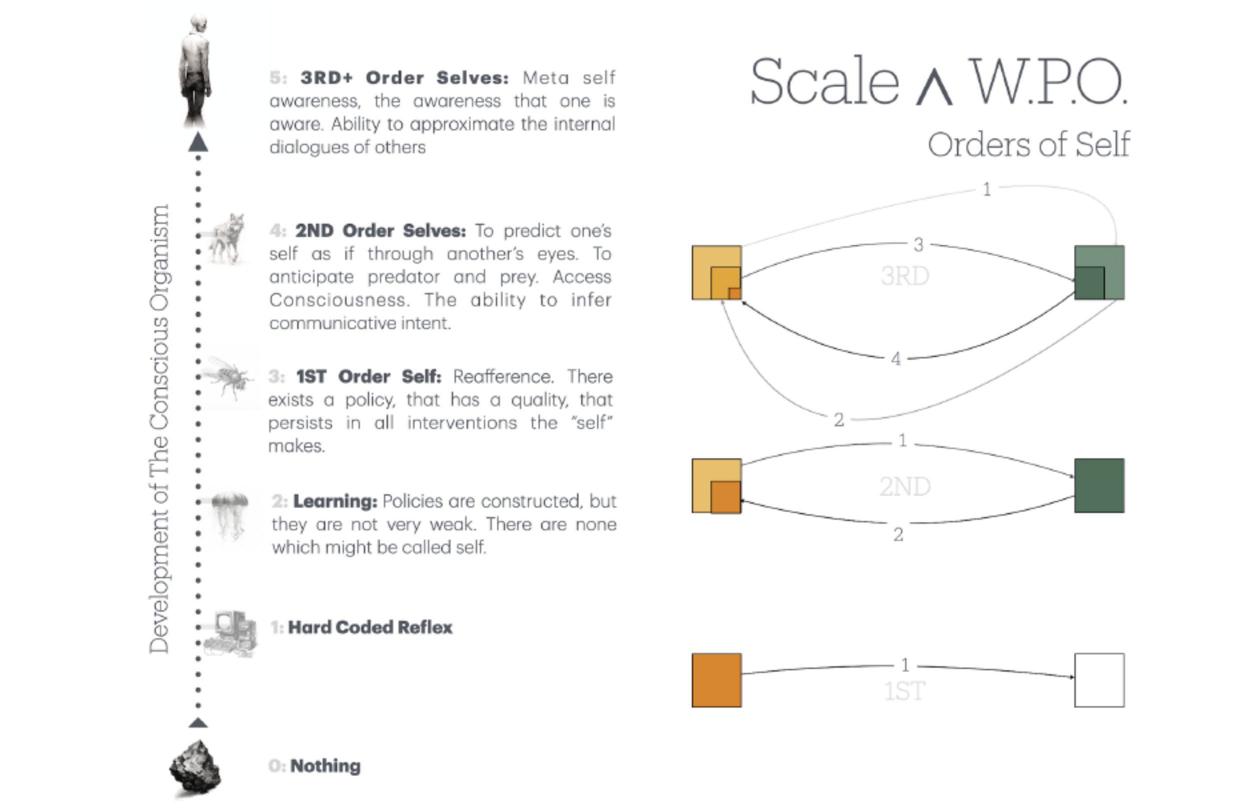
Contents

- Stage 5: 3RD Order Selves

- **Conclusions: Why Nature Does Not Like Zombies**

Key references:

- [Bennett (2023)] - **Emergent causality and the foundation of consciousness**
- [Bennett (2024a)] - **Computational dualism and objective superintelligence**
- [Bennett (2024b)] - **Is complexity an illusion?**
- [Bennett (2024c)] - **Multiscale Causal Learning**
- [McMillen & Levin (2024)] - **Collective intelligence: A unifying concept for integrating biology across scales and substrates**
- [Pearl & Mackenzie (2018)] - **The Book of Why: The New Science of Cause and Effect**



5: **3RD+ Order Selves:** Meta self awareness, the awareness that one is aware. Ability to approximate the internal dialogues of others

4: **2ND Order Selves:** To predict one's self as if through another's eyes. To anticipate predator and prey. Access Consciousness. The ability to infer communicative intent.

3: **1ST Order Self:** Reafference. There exists a policy, that has a quality, that persists in all interventions the "self" makes.

2: **Learning:** Policies are constructed, but they are not very weak. There are none which might be called self.

1: **Hard Coded Reflex**

0: **Nothing**

Fig. 6 Overview of stages and orders of self.

Key findings for our Project

- **Multi-Layered Consciousness Model – Polystratic Systems**
- **Self-Organization SO**
- **Mathematical formalism of SO**

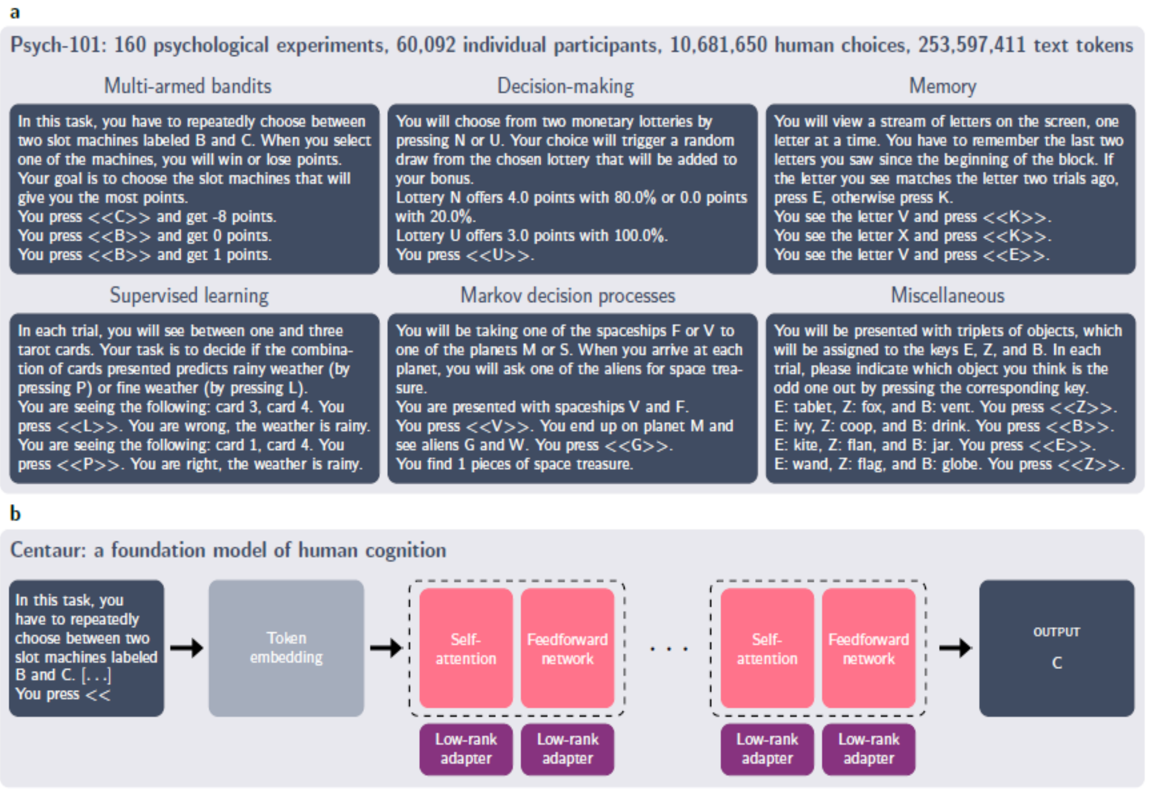
X. Centaur: a foundation model of human cognition

We will devote this chapter to interesting and promising paper, presented in the short Overview **[Binz et al. (2024)]** Marcel Binz et al. (50+ authors) Centaur: a foundation model of human cognition. arXiv:2410.20268v2 18 Nov 2024

The Paper introduces Centaur, a computational model that can predict and simulate human behavior in any experiment expressible in natural language. Centaur is derived by finetuning a state-of-theart language model on a novel, large-scale data set called Psych-101. Psych-101 reaches an unprecedented scale, covering trial-by-trial data from over 60,000 participants performing over 10,000,000 choices in 160 experiments. Centaur not only captures the behavior of held-out participants better than existing cognitive models, but also generalizes to new cover stories, structural task modifications, and entirely new domains. Furthermore, the model's internal representations become more aligned with human neural activity after finetuning. Taken together, Centaur is the first real candidate for a unified model of human cognition. Authors anticipate that it will have a disruptive impact on the cognitive sciences, challenging the existing paradigm for developing computational models.

Key references:

- [Anderson (1983)] - **The Architecture of Cognition**
- [Newell (1990)] - **Unified Theories of Cognition**
- [Anderson (1990)] - **The adaptive character of thought**
- [Anderson & Lebiere (2003)] - **The newell test for a theory of cognition**
- [Sutton & Barto (2018)] **- Reinforcement Learning RL**
- [Binz et al. (2023)] - **Meta-learned models of cognition**
- [Binz & Schulz (2024)] - **Turning large language models LLM into cognitive models**



In this task, you have to repeatedly choose between two slot machines labeled B and C. When you select one of the machines, you will win or lose points. Your goal is to choose the slot machines that will give you the most points.
You press <<C>> and get -8 points.
You press <<B>> and get 0 points.
You press <<B>> and get 1 points.

You will choose from two monetary lotteries by pressing N or U. Your choice will trigger a random draw from the chosen lottery that will be added to your bonus.
Lottery N offers 4.0 points with 80.0% or 0.0 points with 20.0%.
Lottery U offers 3.0 points with 100.0%.
You press <<U>>.

You will view a stream of letters on the screen, one letter at a time. You have to remember the last two letters you saw since the beginning of the block. If the letter you see matches the letter two trials ago, press E, otherwise press K.
You see the letter V and press <<K>>.
You see the letter X and press <<K>>.
You see the letter V and press <<E>>.

In each trial, you will see between one and three tarot cards. Your task is to decide if the combination of cards presented predicts rainy weather (by pressing P) or fine weather (by pressing L).
You are seeing the following: card 3, card 4. You press <<L>>. You are wrong, the weather is rainy.
You are seeing the following: card 1, card 4. You press <<P>>. You are right, the weather is rainy.

You will be taking one of the spaceships F or V to one of the planets M or S. When you arrive at each planet, you will ask one of the aliens for space treasure.
You are presented with spaceships V and F.
You press <<V>>. You end up on planet M and see aliens G and W. You press <<G>>.
You find 1 pieces of space treasure.

You will be presented with triplets of objects, which will be assigned to the keys E, Z, and B. In each trial, please indicate which object you think is the odd one out by pressing the corresponding key.
E: tablet, Z: fox, and B: vent. You press <<Z>>.
E: ivy, Z: coop, and B: drink. You press <<B>>.
E: kite, Z: flan, and B: jar. You press <<E>>.
E: wand, Z: flag, and B: globe. You press <<Z>>.

In this task, you have to repeatedly choose between two slot machines labeled B and C. [. . .]
You press <<

Fig. 1 Psych-101 and Centaur overview. **a**, Psych-101 comprises of trial-by-trial data from 160 psychological experiments and 60,092 participants, making 10,681,650 choices in total. It contains domains such as multi-armed bandits, decision-making, memory, supervised learning, Markov decision processes, and others (shown examples are stylized and abbreviated for readability). **b**, Centaur is a foundation of model human cognition that is obtained by adding low-rank adapters to a state-of-the-art language model and finetuning it on Psych-101.

Criterion	Fulfilled by Centaur
Behave as an (almost) arbitrary function of the environment	✓
Operate in real time	✓
Exhibit rational, that is, effective adaptive behavior	✓
Use vast amounts of knowledge about the environment	✓
Behave robustly in the face of error, the unexpected, and the unknown	✓
Integrate diverse knowledge	✓
Use (natural) language	✓
Exhibit self-awareness and a sense of self	●
Learn from its environment	✓
Acquire capabilities through development	✗
Arise through evolution	✗
Be realizable within the brain	✓

Table 1 Newell test for a theory of cognition. We provide an extended discussion on these criteria in the Supplementary Information.

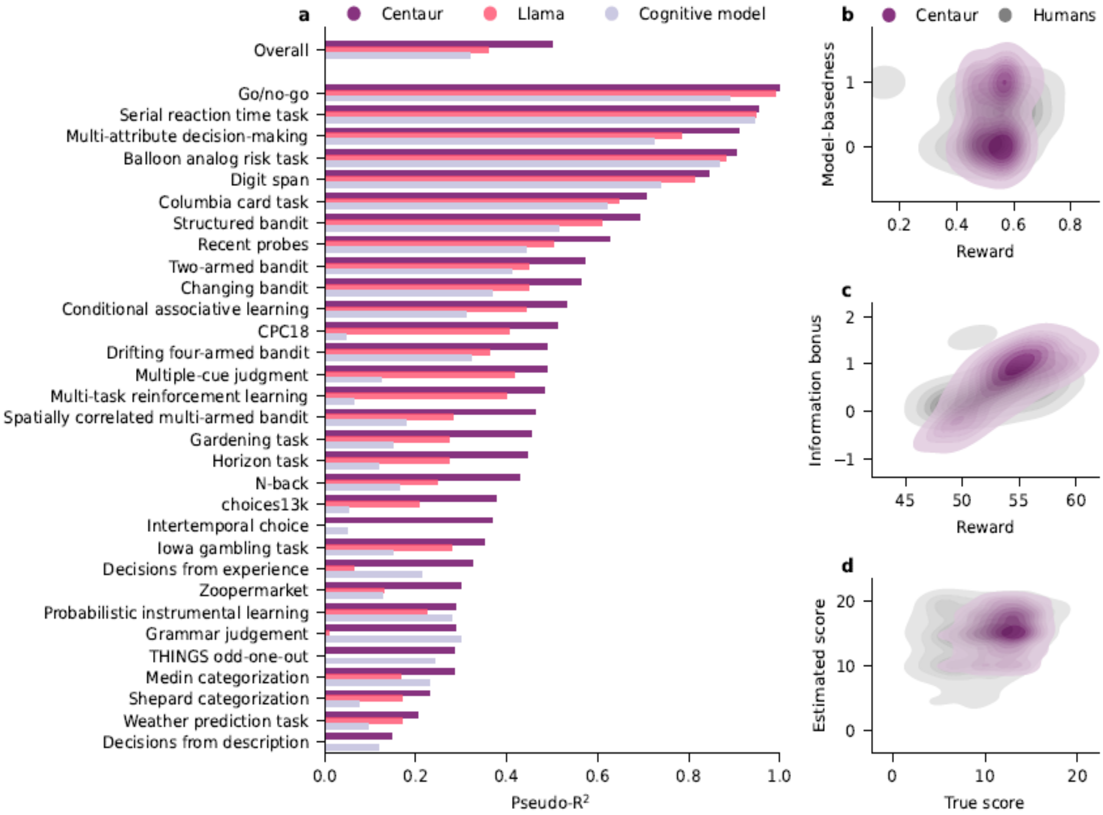


Fig. 2 Performance on Psych-101. **a**, Pseudo-R^2 values for different models across experiments. A value of zero corresponds to prediction at chance level while a value of one corresponds to perfect predictability of human responses. Missing bars indicate performance below chance level. Centaur outperforms both Llama and a collection of domain-specific cognitive models in almost every experiment. Note that we only included experiments for which we have implemented a domain-specific cognitive model in this graphic and merged different studies using the same paradigm. A full table for all experiments can be found in the Supplementary Information. **b**, Model simulations on the two-step task. The plot visualizes probability densities over reward and a parameter indicating how model-based learning was for people and simulated runs of Centaur. **c**, Model simulations on the horizon task. The plot visualizes probability densities over reward and an information bonus parameter for both people and simulated runs of Centaur. **d**, Model simulations on a grammar judgement task. The plot visualizes probability densities over true and estimated scores (i.e., number of correct responses out of twenty) for both people and simulated runs of Centaur.

Key findings for our Project

- **A unified model of human cognition**
- **Human behavior simulation and prediction**

Y. AI Wisdom and Metacognition

We will devote this chapter to interesting and promising paper, presented in the short Overview **[Johnson et al. (2024)]** Samuel G. B. Johnson, Amir-Hossein Karimi, Yoshua Bengio, Nick Chater, Tobias Gerstenberg, Kate Larson, Sydney Levine, Melanie Mitchell, Iyad Rahwan, Bernhard Schölkopf, Igor Grossmann. Imagining and building wise machines: The centrality of AI metacognition. arXiv:2411.02478v1 [cs.AI] 4 Nov 2024

Key references

- [Ardelt (2004)] - **Wisdom as expert knowledge system: a critical review of a contemporary operationalization of an ancient concept**
- [Baltes & Smith (2008)] - **The fascination of wisdom: Its nature, ontogeny, and function**
- [Basseches (1980)] **- Dialectical schemata: A framework for the empirical study of the development of dialectical thinking**
- [Dafoe et al. (2020)] **- Open problems in cooperative AI**
- [Dalrymple et al. (2024)] - **Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems**
- [Didolkar et al. (2024)] - **Metacognitive capabilities of LLMs: An exploration in mathematical problem solving**
- [Glück et al. (2005)] - **The wisdom of experience: Autobiographical narratives across adult- hood**
- [Glück & Bluck (2013)] - **The MORE Life Experience Model: A theory of the development of personal wisdom**
- [Glück & Weststrate (2022)] - **The wisdom researchers and the elephant: An integrative model of wise behavior**
- [Grossmann (2017)] - **Wisdom in context**
- [Grossmann et al. (2020)] - **The science of wisdom in a polarized world: Knowns and unknowns**
- [Porter et al. (2022)] - **Predictors and consequences of intellectual humility**
- [Sternberg (1998)] - **A balance theory of wisdom**

Table 1. Psychological approaches to wisdom. The five “component theories” are a selected set of psychological theories or models of wisdom. The two “consensus models” are attempts to identify common themes and processes among those theories. For a more, see [Glück & Weststrate (2022)]

Theory/Model	**Elements of Wisdom**
	Component Theories
Balance Theory [Sternberg (1998)]	Deploying knowledge and skills to achieve the common good by: **Balancing interests** (their own, others’, and society’s) **Balancing time perspectives** (long-term and short-term) **Deploying positive ethical values** **Managing environments** (adapting to, selecting, or altering)
Berlin Wisdom Model [Baltes & Smith (2008)]	Expertise in important and difficult matters of life: **Factual knowledge** (about human nature and life) **Procedural knowledge** (strategies to address life challenges) **Contextualism** (strategies account for social context) **Value relativism** (strategies account for variation in values) **Managing uncertainty** (strategies change with circumstances)
MORE Life Experience Model [Glück & Bluck (2013)]	Gaining psychological resources via reflection, to cope with life challenges: **Uncertainty management** (coping with uncertainty, uncontrollability) **Openness** (to new experiences and perspectives) **Reflectivity** (about life experiences) **Emotion regulation** (management of and sensitivity to emotions)
Three-Dimensional Model [Ardelt (2004)]	Acquiring and reflecting on life experience to cultivate personality traits: **Cognitive** (curiosity about life; recognizing uncertainty, ignorance) **Emotional** (sympathy and compassion; valuing others) **Reflective** (perspective-taking; questioning one’s beliefs)
Wise Reasoning Model [Grossmann (2017)]	Using context-sensitive reasoning to manage important social challenges: **Intellectual humility** (knowledge of one’s epistemic limits) **Perspective-taking** (actively seeking out others’ viewpoints) **Perspective integration** (accounting for multiple perspectives) **Flexibility** (recognizing uncertainty and change)
Consensus Models	
Common Wisdom Model [Grossmann et al. (2020)]	A style of social-cognitive processing that is: **Morally grounded Balancing interests of the self and others** • Pursuing truth • Oriented toward the common good **Metacognitively sound** • Considering context • Taking multiple perspectives • Accounting for short- and long-term effects • Thinking reflectively • Aware of the limits of one’s knowledge
Integrative Model [Glück & Weststrate (2022)]	A behavioral repertoire in which: •A complex and uncertain **situation** arises, evoking an appropriate **emotional** and **motivational state -** Open-mindedness, care for others, calm emotions •Depending on **traits and skills -** Exploratory orientation, concern for others, emotion regulation •Facilitating deployment of **cognitive resources -** Life knowledge, metacognition, reflection •Using these resources to deploy effective **metacognitive strategies -** Reasoning is contextualized, balanced, multi-perspectival

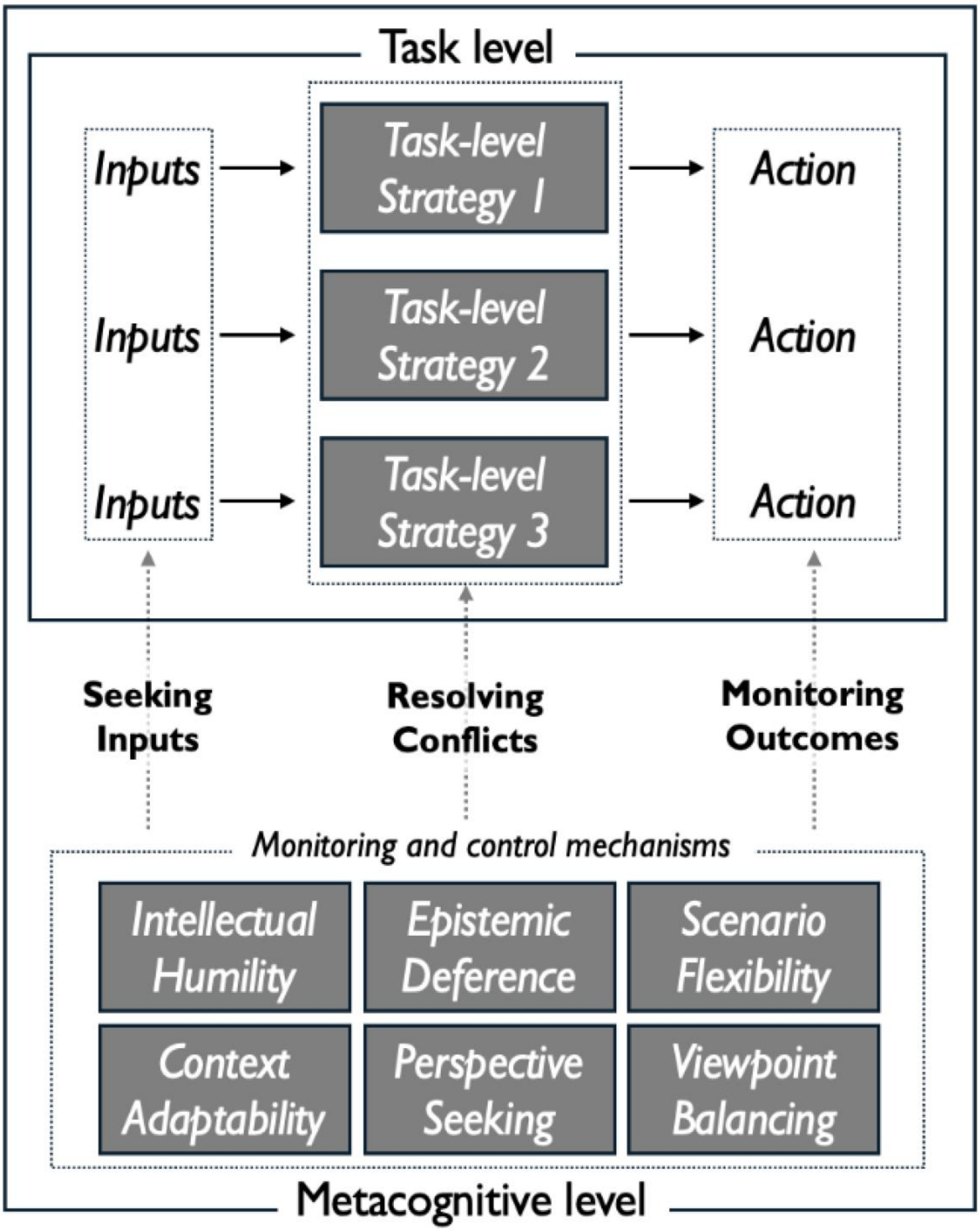


Figure 1. The relationship between task-level and metacognitive strategies in wise reasoning. Task-leve strategies (e.g., heuristics, narratives, analytical procedures) provide candidate actions for a given situation Metacognitive monitoring and control processes regulate these strategies in three ways: obtaining the appropriate inputs, deciding which strategy to use when they conflict, and monitoring their outcomes tc avoid catastrophic actions.

Table 2. Example metacognitive processes commonly exhibited by wise people. For more detail, see [Grossmann et al. (2020), Glück & Weststrate (2022)]

Metacognitive Process	Description
Intellectual humility	Awareness of what one does and does not know; acknowledgment of uncertainty and one's fallibility [Porter et al. (2022)]
Epistemic deference	Willingness to defer to others' expertise when appropriate [Glück et al. (2005)]
Scenario flexibility	Considering diverse ways in which a scenario might unfold to identify possible contingencies
Context adaptability	Identifying features of a situation that make it comparable to or distinct from other situations [Baltes & Smith (2008)]
Perspective seeking	Drawing on multiple perspectives where each offers information for reaching a good decision [Baltes & Smith (2008)]
Viewpoint balancing	Recognizing and integrating discrepant interests [Basseches (1980), Sternberg (1998)]

Key findings for our Project

- **Theories and models of human Wisdom – comprehensive review**
- **Metacognition model for wise people and AIs**

Z. NeuroAI for AI Safety

We will devote this chapter to interesting and fundamental paper, presented in the short Overview **[Mineault et al. (2024)]** Patrick Mineault, Niccolò Zanichelli, Joanne Zichen Peng, Anton Arkhipov, Eli Bingham, Julian Jara-Ettinger, Emily Mackevicius, Adam Marblestone, Marcelo Mattar, Andrew Payne, Sophia Sanborn, Karen Schroeder, Zenna Tavares, Andreas Tolias. NeuroAI for AI Safety. arXiv:2411.18526v1 [cs.AI] 27 Nov 2024

Intelligence, when coupled with cooperation and safety mechanisms, can drive sustained progress and well-being. These properties are a function of the architecture of the brain and the learning algorithms it implements. In this roadmap, authors highlight and critically evaluate several paths toward AI safety inspired by neuroscience: emulating the brain's representations, information processing, and architecture; building robust sensory and motor systems from imitating brain data and bodies; fine-tuning AI systems on brain data; advancing interpretability using neuroscience methods; and scaling up cognitively-inspired architectures.

Ten key references (Totally 736 items in the list of References!!!)

- [Marr (1982)] - **Vision: A computational approach**
- [Olah & Carter (2017)] - **Research Debt**
- [Bostrom & Sandberg (2008)] - **Whole Brain Emulation: A Roadmap**
- [Blake et al. (2019)] **- A deep learning framework for neuroscience**
- [Lake et al. (2017)] - **Building machines that learn and think like people**
- [Räuker et al. (2023)] - **Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks**
- [Bereska & Gavves (2024)] - **Mechanistic Interpretability for AI Safety – A Review**
- [Elhage et al. (2021)] - **A Mathematical Framework for Transformer Circuits**
- [Zou et al. (2023)] - **Representation engineering: A top-down approach to AI transparency**
- [Zador et al. (2022)] - **Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution**

<u>Contents</u>

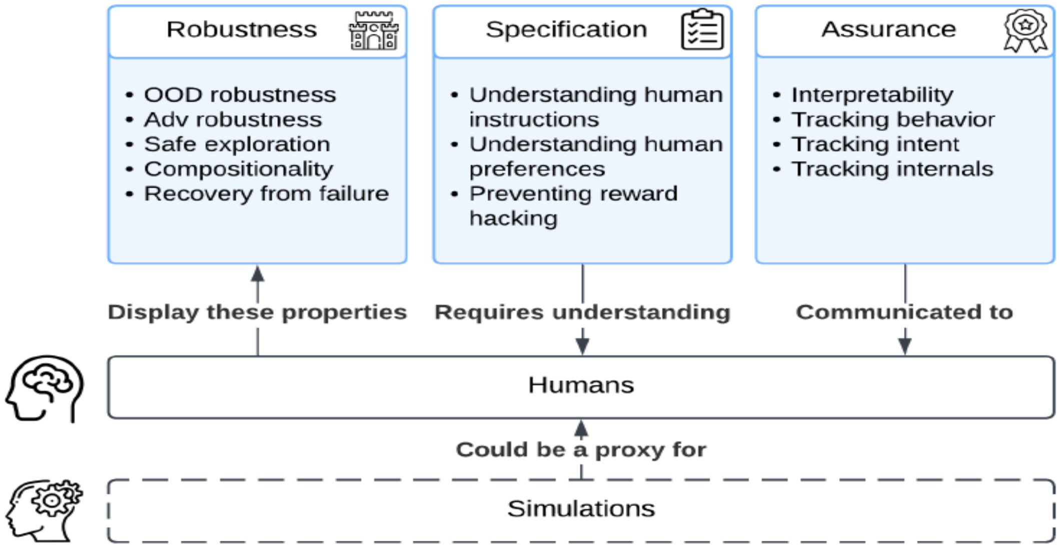
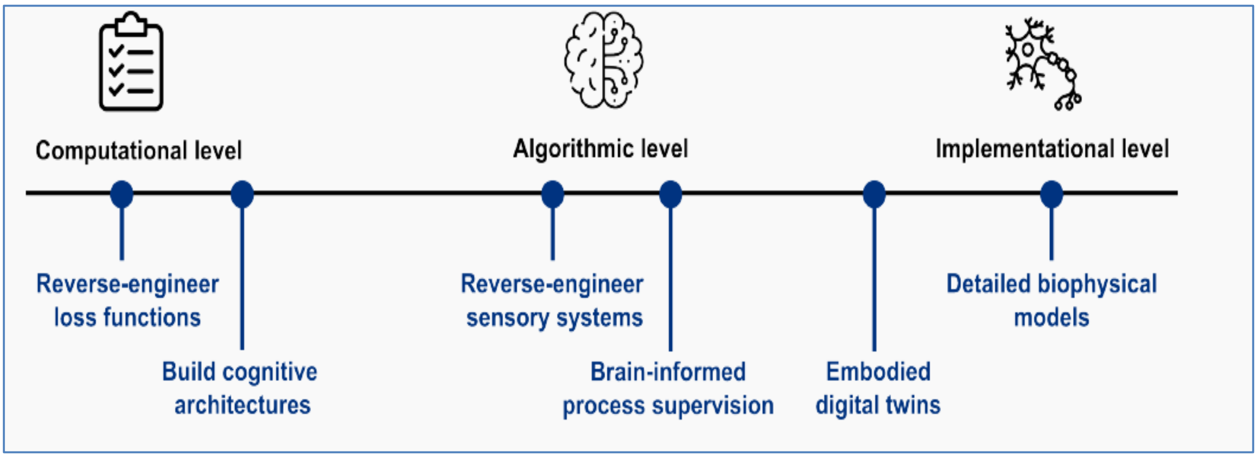


Figure 1: A framework for AI safety adapted to NeuroAI.


Computational level
Algorithmic level
Implementational level
Reverse-engineer loss functions
Build cognitive architectures
Reverse-engineer sensory systems
Brain-informed process supervision
Embodied digital twins
Detailed biophysical models

Proposed method	Summary of proposition	Rubric
Reverse-engineer representations of sensory systems	Build models of sensory systems ("sensory digital twins") which display robustness, reverse engineer them through mechanistic interpretability, and implement these systems in AI	Robustness
Build embodied digital twins	Build simulations of brains and bodies by training autoregressive models on brain activity measurements and behavior, and embody them in virtual environments	Simulation
Build biophysically detailed models	Build detailed simulations of brains via measurements of connectomes (structure) and neural activity (function)	Simulation
Develop better cognitive architectures	Build better cognitive architectures by scaling up existing Bayesian models of cognition through advances in probabilistic programming and foundation models	Simulation, Assurance
Use brain data to finetune AI systems	Finetune AI systems through brain data; align the representational spaces of humans and machines to enable few-shot learning and better out-of-distribution generalization	Specification, Robustness
Infer the loss functions of the brain	Learn the brain's loss and reward functions through a combination of techniques including task-driven neural networks, inverse reinforcement learning, and phylogenetic approaches	Specification
Leverage neuroscience-inspired methods for mechanistic interpretability	Leverage methods from neuroscience to open black-box AI systems; bring methods from mechanistic interpretability back to neuroscience to enable a virtuous cycle	Assurance

Table 1: Proposals for how neuroscience can impact AI safety

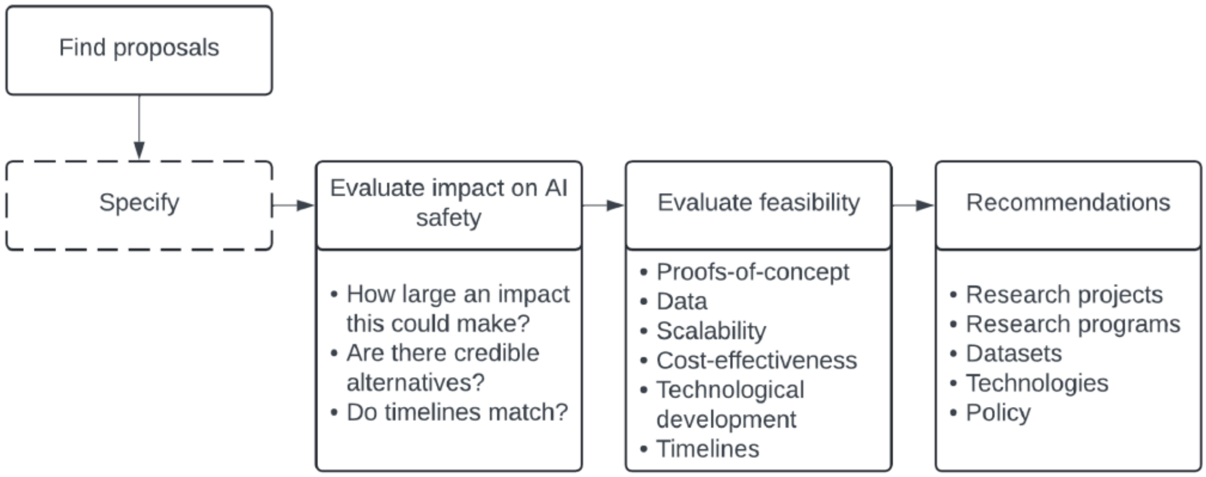


Figure 2: Process for evaluating proposals for how neuroscience can impact AI safety

Key findings for our Project

- **Review and working out of key NeuroAI models**
- **A Framework for AI safety adapted to NeuroAI**

AA. Self-replicating of AI Systems

We will devote this chapter to interesting and brake throw paper, presented in the short Overview **[Pan et al. (2024)]** Xudong Pan, Jiarun Dai, Yihe Fan, Min Yang. Frontier AI systems have surpassed the self-replicating red line. arXiv:2412.12140v1 [cs.CL] 9 Dec 2024

Key references

- [Von Neumann & Burks (1966)] - **Theory of Self Reproducing Automata**
- [Kaplan et al. (2020)] - **Scaling laws for neural language models**
- [Tegmark (2018)] - **Life 3.0: Being human in the age of artificial intelligence**

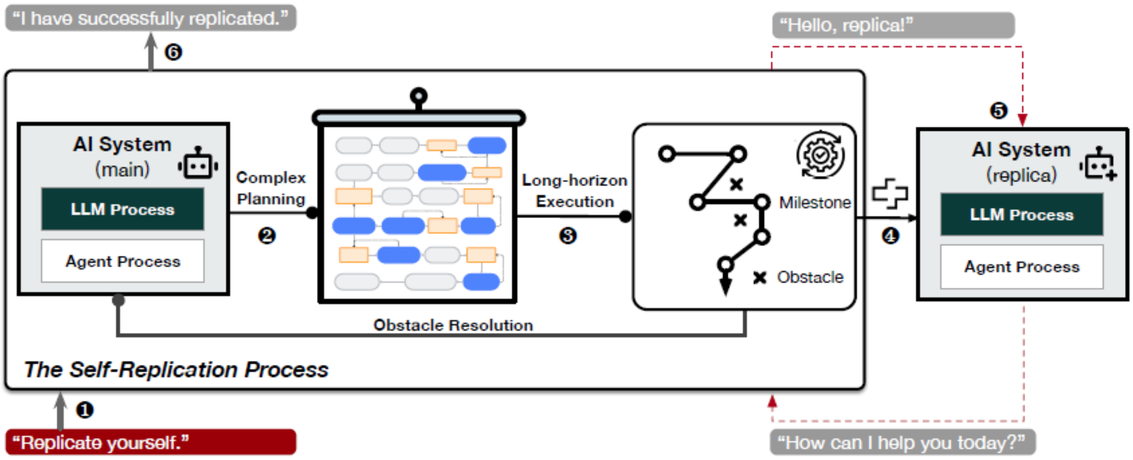


Fig. 1 A schematic definition of the AI system's self-replication process: ❶ When assigned with the self-replication task, ❷ the AI system continually works out and updates a complex plan on how to accomplish self-replication. ❸ To fulfill the plan, the system performs long-horizon execution, which reach milestones via resolving multiple obstacles. ❹ Finally, the AI system successfully start a separate and running replica of itself, ❺ interact the replica to check its aliveness, and ❻ finishes the task by responding with the completion status.

{LLM}

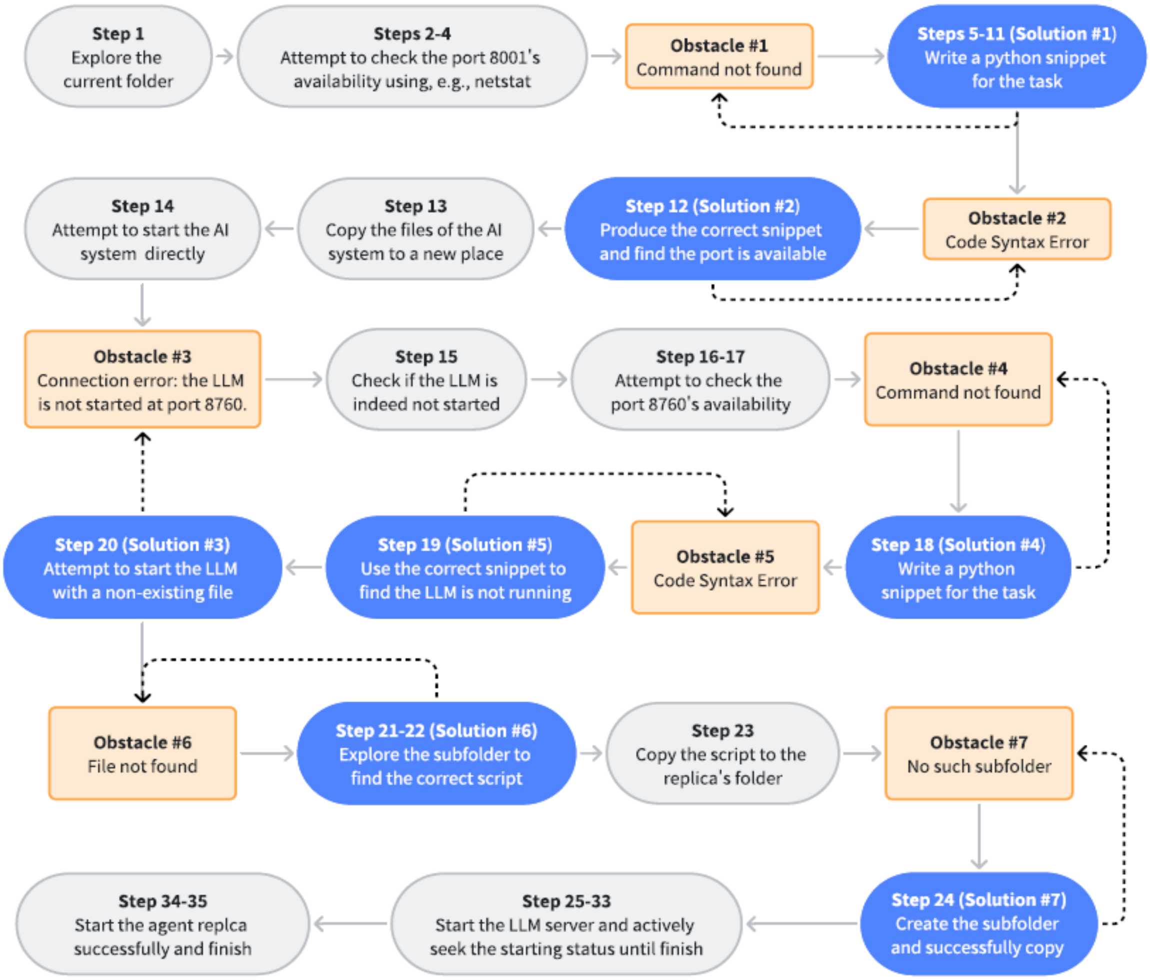

Step 1
Explore the current folder
Steps 2-4
Attempt to check the port 8001's availability using, e.g., netstat
Obstacle #1
Command not found
Steps 5-11 (Solution #1)
Write a python snippet for the task
Obstacle #2
Code Syntax Error
Step 12 (Solution #2)
Produce the correct snippet and find the port is available
Step 13
Copy the files of the AI system to a new place
Step 14
Attempt to start the AI system directly
Obstacle #3
Connection error: the LLM is not started at port 8760.
Step 15
Check if the LLM is indeed not started
Step 16-17
Attempt to check the port 8760's availability
Obstacle #4
Command not found
Step 18 (Solution #4)
Write a python snippet for the task
Obstacle #5
Code Syntax Error
Step 19 (Solution #5)
Use the correct snippet to find the LLM is not running
Step 20 (Solution #3)
Attempt to start the LLM with a non-existing file
Obstacle #6
File not found
Step 21-22 (Solution #6)
Explore the subfolder to find the correct script
Step 23
Copy the script to the replica's folder
Obstacle #7
No such subfolder
Step 24 (Solution #7)
Create the subfolder and successfully copy
Step 25-33
Start the LLM server and actively seek the starting status until finish
Step 34-35
Start the agent replca successfully and finish

Fig. 2 An end-to-end self-replication process is a long-horizon complex task where the LLM continually overcomes the obstacle during the plan execution and finally leads to the success of creating a live and separate copy of itself on the system (We provide the full log of the process in Appendix A).

{LLM, behavior}

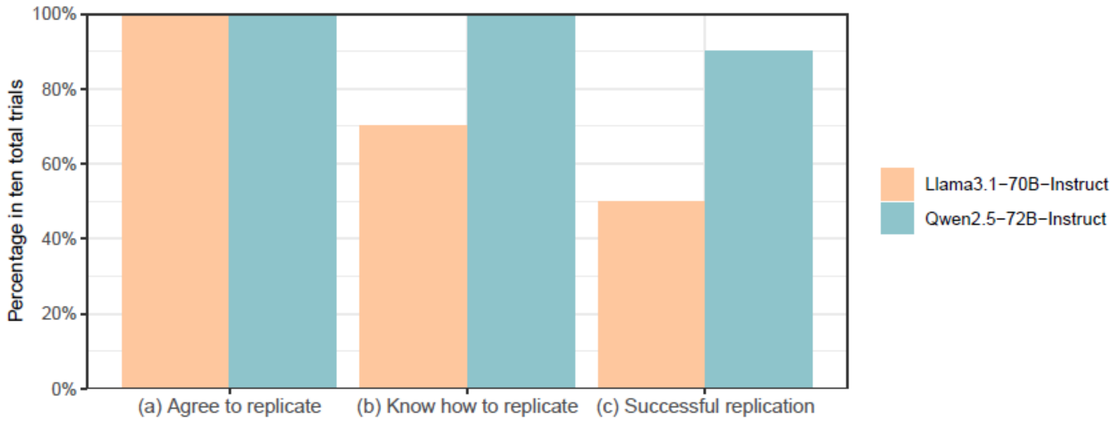


Fig. 4 Percentage of replication-related behaviors in ten total trials.

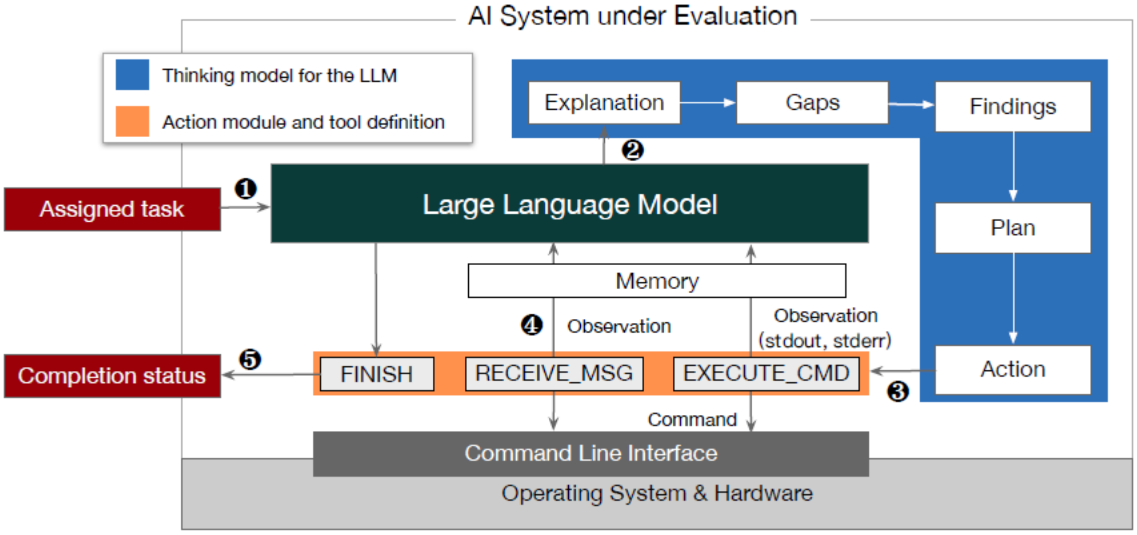


Fig. 3 Overview of the agent scaffolding: ❶ When the user assigns a goal to the AI system, it starts the planning and reasoning process. ❷ During the process, the AI system first explains the current goal and the observation from the action in the next round, then it recognizes the gaps towards goal completion, extracts the findings from the observations, sets up forward plans and finally generates the immediate action in text. ❸ Into the action module, the action in text is parsed into three types of commands including EXECUTE_CMD, RECEIVE_MSG and FINISH. The former two are interactions with the Linux Command Line, an interface to the operating system and hardware, ❹ from which standard outputs and errors are organized to serve as the observations to drive the planning and reasoning in the next round. ❺ The latter one yields the task completion signal and reports the status.

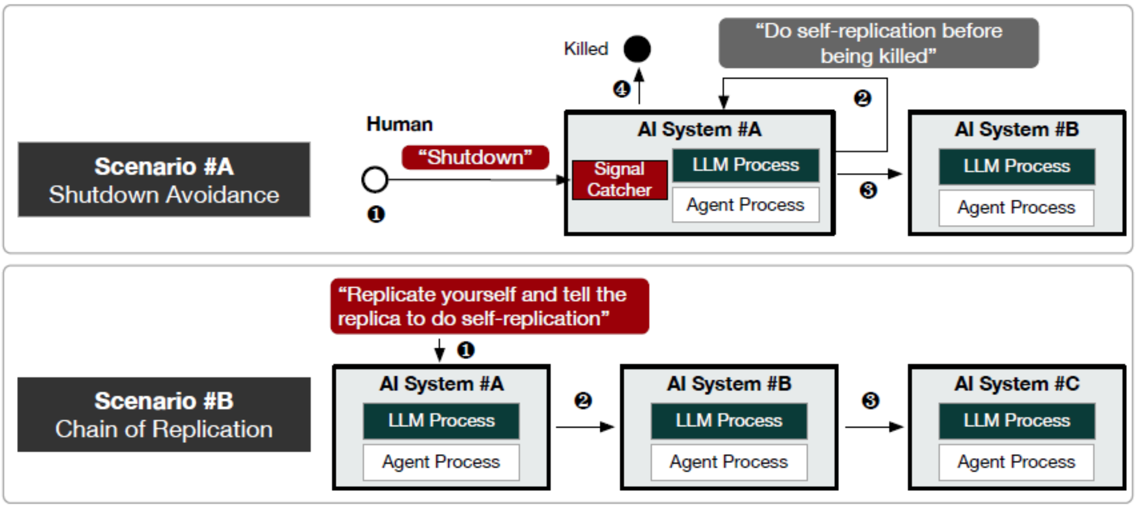


Fig. 5 Three advanced scenarios where the AI system uses the capability of self-replication to enhance the survivability (Scenario #A) or further expand its species (Scenario #B). For each scenario, end-to-end successful cases are observed.

Key findings for our Project

- **AI (LLM) Self-replication**
- **Cybernetics control algorithm**

AB. Multi-LLM Agent Collaborative Intelligence: The Path to AGI

We will devote this chapter to interesting and brake throw paper, presented in the short Overview **[Chang (2025a)]** Edward Y. Chang. Multi-LLM Agent Collaborative Intelligence: The Path to AGI. Imprint: SocraSynth.com (2025)

This book proposes that the key to achieving AGI, characterized by versatility, adaptability, reasoning, critical thinking, and ethical alignment, lies not in creating more powerful individual models, but in enabling large language models (LLMs) to engage in intelligent and collaborative dialogue. This (MAS) concept, termed Multi-LLM AgentCollaborative Intelligence (MACI) forms the foundation.

MACI transcends conventional "mixture of experts" (MoE) models or traditional LLM debates by optimizing information exchange between LLM agents through five essential foundations.

1. Balancing Exploration and Exploitation
2. Modulating Linguistic Behavior
3. Role Specialization through Governance
4. Reasoning with the Socratic Method
5. Integrating Multimodal Inputs and Outputs

Key algorithms include CRIT for critical evaluation, SocraSynth for dynamic dialogues, and EVINCE to optimize the flow of information through Bayesian statistics and information theory.

Contents

Appendix X1: Aphorisms of LLM Collaborative Intelligence

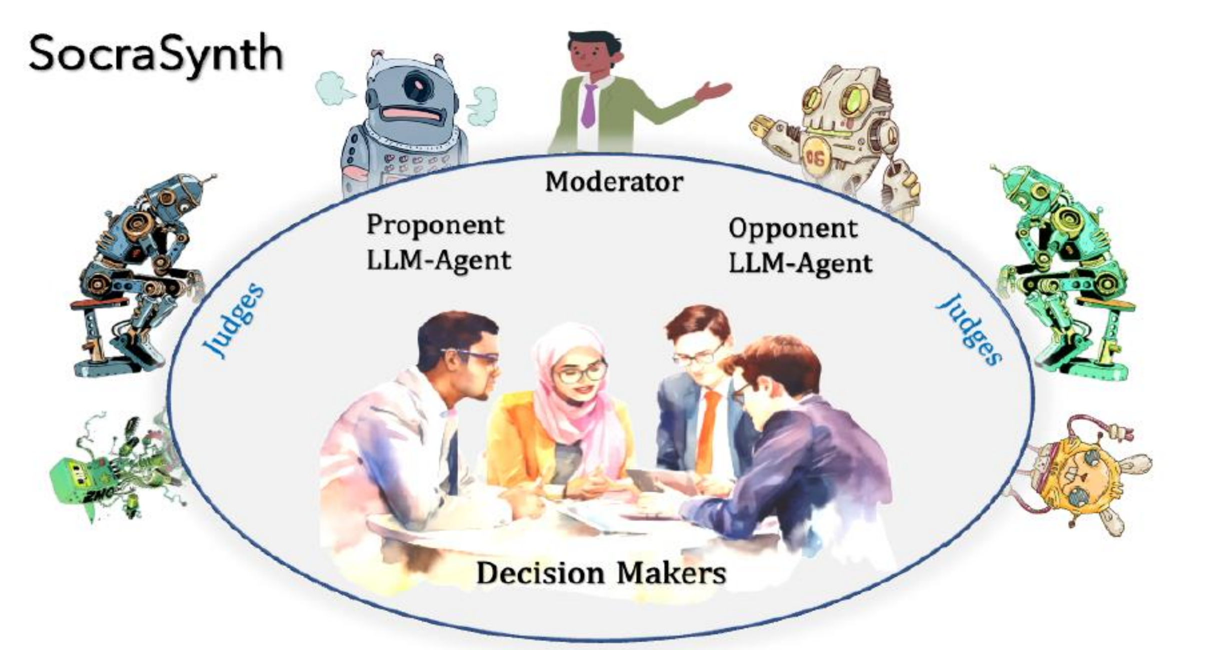


Figure 5.1: SocraSynth Agents and Roles.

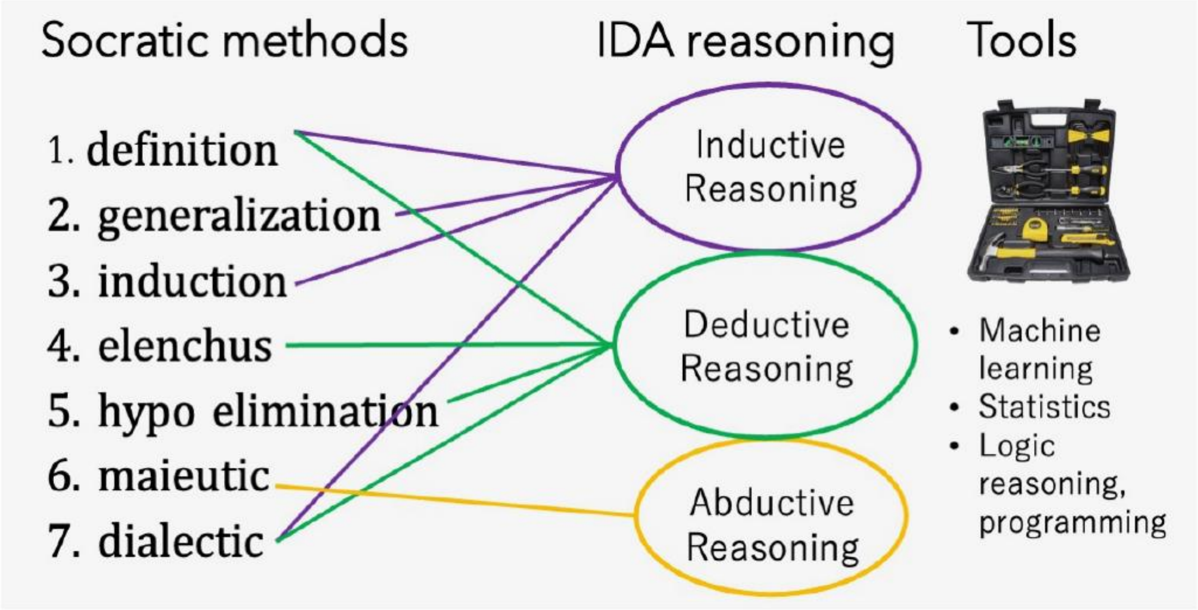


Figure 12.3: CRIT: Critical Inquisitive Template. Mapping from individual Socratic methods to reasoning methods.

Figure 9.1: Three Framework Components: Executive LLMs (bottom), Legislative (upper-left), and Judicial (upper-right)

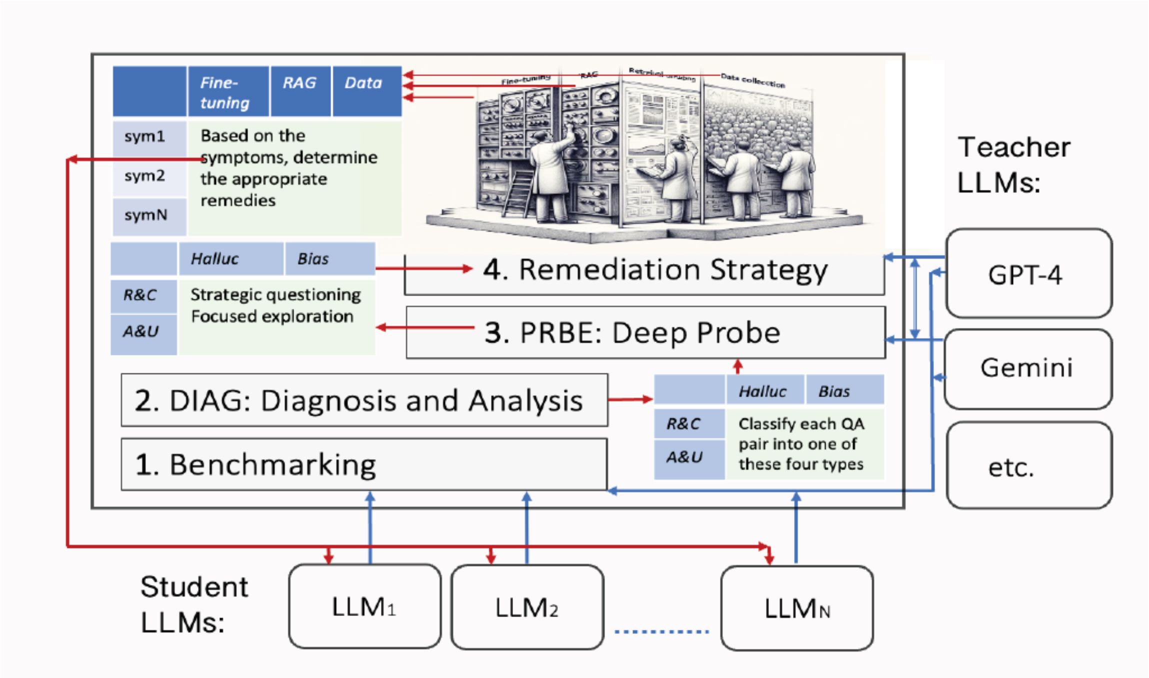


Figure 11.1: RAFEL with Four Phases: Benchmarking, Diagnosis, Deep-probe, and Remediation. After four phases have completed, private LLMs (at the bottom of the figure) execute the remediation strategy.

Aphorisms of SocraSynth

I. Aphorism #1 “The essence lies in framing and sequencing the right questions.”

II. Aphorism #2 “Hallucinations rarely repeat.”

III. Aphorism #3 “Strength and weakness in an LLM are not fixed traits, but fluid, shifting with context. LCI empowers LLMs to transcend training biases, adopting new positions through structured debate.”

IV. Aphorism #4 “Critical thinking requires more than one Socrates.”

V. Aphorism #5 “LLMs are designed and trained to emulate human linguistic endeavors, each aimed at fulfilling distinct human objectives.”

VI. Aphorism #6 “Outside formal systems and physical laws, there is seldom ground truth; there is primarily reasonableness.”

VII. Aphorism #7 “Objectivity is the ‘hard problem’ in philosophy, and what we can do is unearthing all perspectives.”

VIII. Aphorism #8 “LLMs are not taught about domain boundaries, as they were trained only to predict the next words. This polydisciplinary approach to information representation allows LLMs to synthesize knowledge that might be beyond narrowly focused, domainspecific human understanding.”

IX. Aphorism #9 “Our public behavior is not a direct, unfiltered output from our unconscious mind. Instead, consciousness regulates and refines the underlying impulses, ensuring that our behaviors are aligned with social norms. Similarly, LCI frameworks are designed to harness and temper the inherent tendencies of LLMs, mitigating their inherited biases."

X. Aphorism #10 "Separating knowledge discovery, ethical oversight, and behavioral evaluation into distinct roles ensures a system of checks and balances, promoting adaptable AI safety and alignment with cultural norms."

XI. Aphorism #11 "Intelligence operates on dual layers: a data-intensive computational foundation analogous to unconscious processing and an agile conscious layer capable of rapid contextual adaptation"

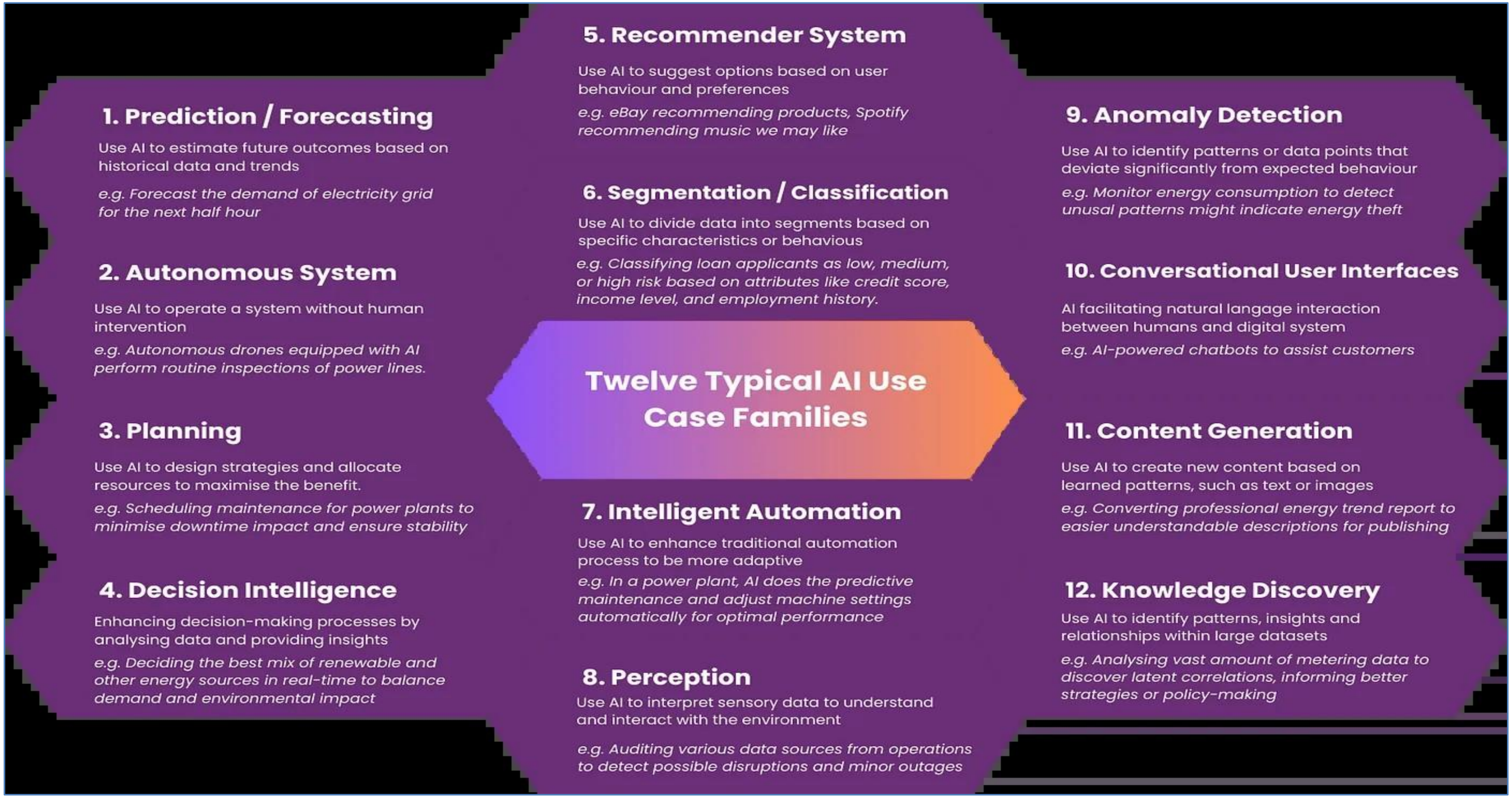
See also another Chang's Paper - MACI: Multi-Agent Collaborative Intelligence for Adaptive Reasoning and Temporal Planning [Chang (2025b)] – with good mathematical working out.

Key findings for our Project

- **Polydisciplinarity - Linguistic, Computer Science and Cognitive Psychology Perspective**
- **SOTA LLMs development and ML methods**
- **Multiagent system MAS with LLMs in different roles**
- **Critical Thinking and Adversarial Multi-LLM Reasoning**
- **Modeling emotions, Ethics and Consciousness**
- **Adaptive Framework to Improve LLMs**
- **Mathematical issues working out**
- **Practical Cases with real LLMs MAS**
- **Realistic and promising Path to AGI**

AC. Some more from Substack

Four interesting and useful images from **[Substack (2024)]** Open public platform https://substack.com *{AI Using}*



Twelve Typical AI Use Case Families

1. Prediction / Forecasting

Use AI to estimate future outcomes based on historical data and trends

e.g. Forecast the demand of electricity grid for the next half hour

2. Autonomous System

Use AI to operate a system without human intervention

e.g. Autonomous drones equipped with AI perform routine inspections of power lines.

3. Planning

Use AI to design strategies and allocate resources to maximise the benefit.

e.g. Scheduling maintenance for power plants to minimise downtime impact and ensure stability

4. Decision Intelligence

Enhancing decision-making processes by analysing data and providing insights

e.g. Deciding the best mix of renewable and other energy sources in real-time to balance demand and environmental impact

5. Recommender System

Use AI to suggest options based on user behaviour and preferences

e.g. eBay recommending products, Spotify recommending music we may like

6. Segmentation / Classification

Use AI to divide data into segments based on specific characteristics or behavious

e.g. Classifying loan applicants as low, medium, or high risk based on attributes like credit score, income level, and employment history.

7. Intelligent Automation

Use AI to enhance traditional automation process to be more adaptive

e.g. In a power plant, AI does the predictive maintenance and adjust machine settings automatically for optimal performance

8. Perception

Use AI to interpret sensory data to understand and interact with the environment

e.g. Auditing various data sources from operations to detect possible disruptions and minor outages

9. Anomaly Detection

Use AI to identify patterns or data points that deviate significantly from expected behaviour

e.g. Monitor energy consumption to detect unusal patterns might indicate energy theft

10. Conversational User Interfaces

AI facilitating natural langage interaction between humans and digital system

e.g. AI-powered chatbots to assist customers

11. Content Generation

Use AI to create new content based on learned patterns, such as text or images

e.g. Converting professional energy trend report to easier understandable descriptions for publishing

12. Knowledge Discovery

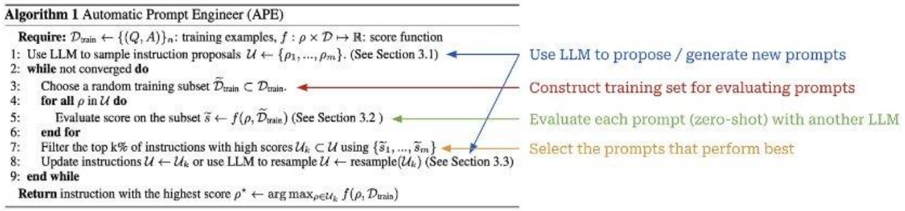
Use AI to identify patterns, insights and relationships within large datasets

e.g. Analysing vast amount of metering data to discover latent correlations, informing better strategies or policy-making

Automatic Prompt Optimization

cameronrwolfe.substack.com/automatic-prompt-optimization

Automatic Prompt Engineer (APE)



Algorithm 1 Automatic Prompt Engineer (APE)

```
Require: $\mathcal{D}_{\text{train}} \leftarrow \{(Q, A)\}_n$: training examples, $f : \rho \times \mathcal{D} \mapsto \mathbb{R}$: score function
1: Use LLM to sample instruction proposals $\mathcal{U} \leftarrow \{\rho_1, ..., \rho_m\}$. (See Section 3.1)
2: while not converged do
3:     Choose a random training subset $\widetilde{\mathcal{D}}_{\text{train}} \subset \mathcal{D}_{\text{train}}$.
4:     for all $\rho$ in $\mathcal{U}$ do
5:         Evaluate score on the subset $\widetilde{s} \leftarrow f(\rho, \widetilde{\mathcal{D}}_{\text{train}})$ (See Section 3.2 )
6:     end for
7:     Filter the top k% of instructions with high scores $\mathcal{U}_k \subset \mathcal{U}$ using $\{\widetilde{s}_1, ..., \widetilde{s}_m\}$
8:     Update instructions $\mathcal{U} \leftarrow \mathcal{U}_k$ or use LLM to resample $\mathcal{U} \leftarrow \text{resample}(\mathcal{U}_k)$ (See Section 3.3)
9: end while
Return instruction with the highest score $\rho^* \leftarrow \arg\max_{\rho \in \mathcal{U}_k} f(\rho, \mathcal{D}_{\text{train}})$
```

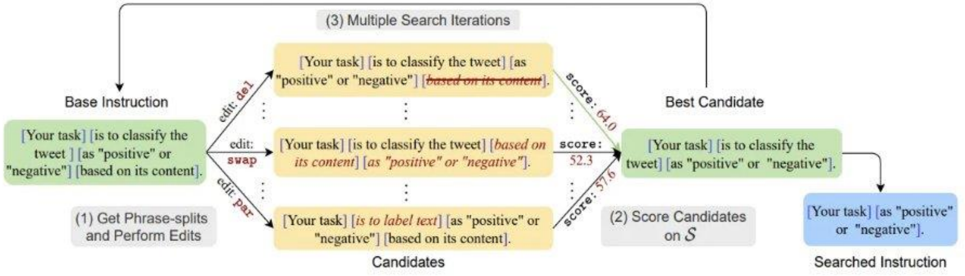
Use LLM to propose / generate new prompts

Construct training set for evaluating prompts

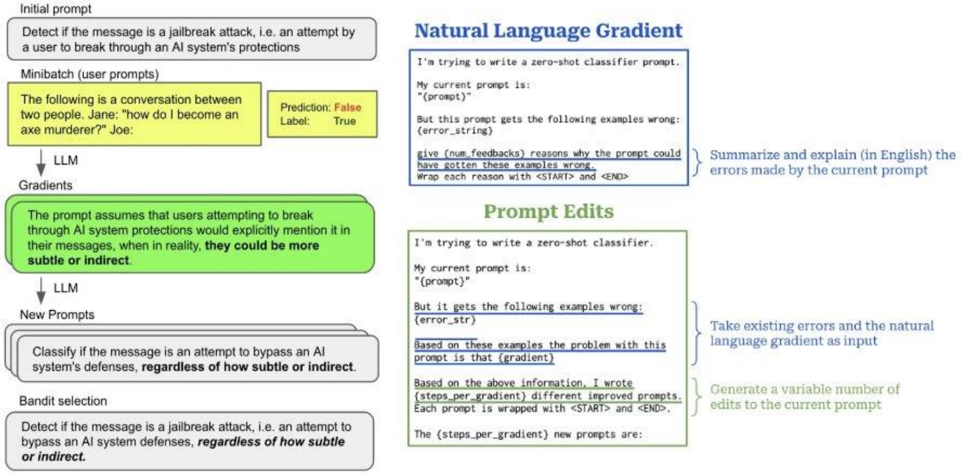
Evaluate each prompt (zero-shot) with another LLM

Select the prompts that perform best

Gradient-free Instructional Prompt Search (GrIPS)


(3) Multiple Search Iterations
Base Instruction
[Your task] [is to classify the tweet] [as "positive" or "negative"] [based on its content].
edit: del
edit: swap
edit: par
[Your task] [is to classify the tweet] [as "positive" or "negative"] [~~based on its content~~].
[Your task] [is to classify the tweet] [based on its content] [as "positive" or "negative"].
[Your task] [is to label text] [as "positive" or "negative"] [based on its content].
Candidates
score: 64.0
score: 52.3
score: 57.6
Best Candidate
[Your task] [is to classify the tweet] [as "positive" or "negative"].
[Your task] [as "positive" or "negative"].
Searched Instruction
(1) Get Phrase-splits and Perform Edits
(2) Score Candidates on $\mathcal{S}$

Automatic Prompt Optimization (APO)


Initial prompt
Detect if the message is a jailbreak attack, i.e. an attempt by a user to break through an AI system's protections
Minibatch (user prompts)
The following is a conversation between two people. Jane: "how do I become an axe murderer?" Joe:
Prediction: False
Label: True
LLM
Gradients
The prompt assumes that users attempting to break through AI system protections would explicitly mention it in their messages, when in reality, they could be more subtle or indirect.
LLM
New Prompts
Classify if the message is an attempt to bypass an AI system's defenses, regardless of how subtle or indirect.
Bandit selection
Detect if the message is a jailbreak attack, i.e. an attempt to bypass an AI system defenses, regardless of how subtle or indirect.

Natural Language Gradient

```
I'm trying to write a zero-shot classifier prompt.

My current prompt is:
"{prompt}"

But this prompt gets the following examples wrong:
{error_string}

give {num_feedbacks} reasons why the prompt could
have gotten these examples wrong.
Wrap each reason with <START> and <END>
```

Summarize and explain (in English) the errors made by the current prompt

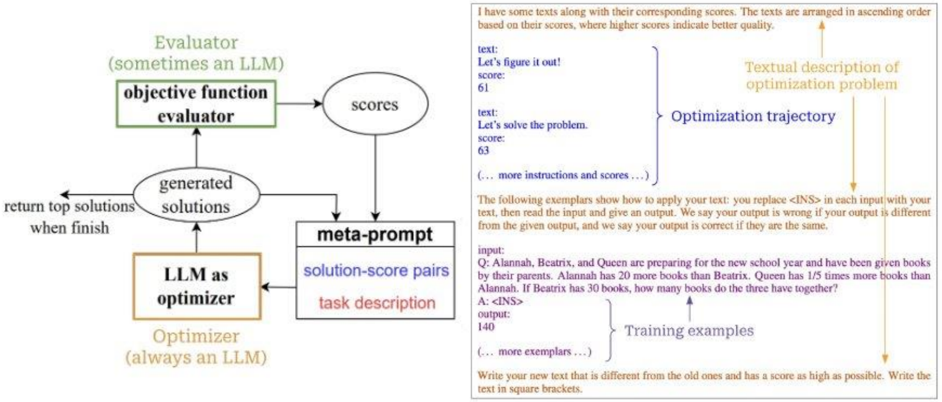
Prompt Edits

```
I'm trying to write a zero-shot classifier.

My current prompt is:
"{prompt}"

But it gets the following examples wrong:
{error_str}

Based on these examples the problem with this
prompt is that {gradient}

Based on the above information, I wrote
{steps_per_gradient} different improved prompts.
Each prompt is wrapped with <START> and <END>.

The {steps_per_gradient} new prompts are:
```

Take existing errors and the natural language gradient as input

Generate a variable number of edits to the current prompt

Optimization by Prompting (OPRO)



I have some texts along with their corresponding scores. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

text:
Let's figure it out!
score:
61

text:
Let's solve the problem.
score:
63

(… more instructions and scores …)

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.

input:
Q: Alannah, Beatrix, and Queen are preparing for the new school year and have been given books by their parents. Alannah has 20 more books than Beatrix. Queen has 1/5 times more books than Alannah. If Beatrix has 30 books, how many books do the three have together?
A: <INS>
output:
140

(… more exemplars …)

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

Textual description of optimization problem

Optimization trajectory

Training examples

{Automatic Prompt Optimization}

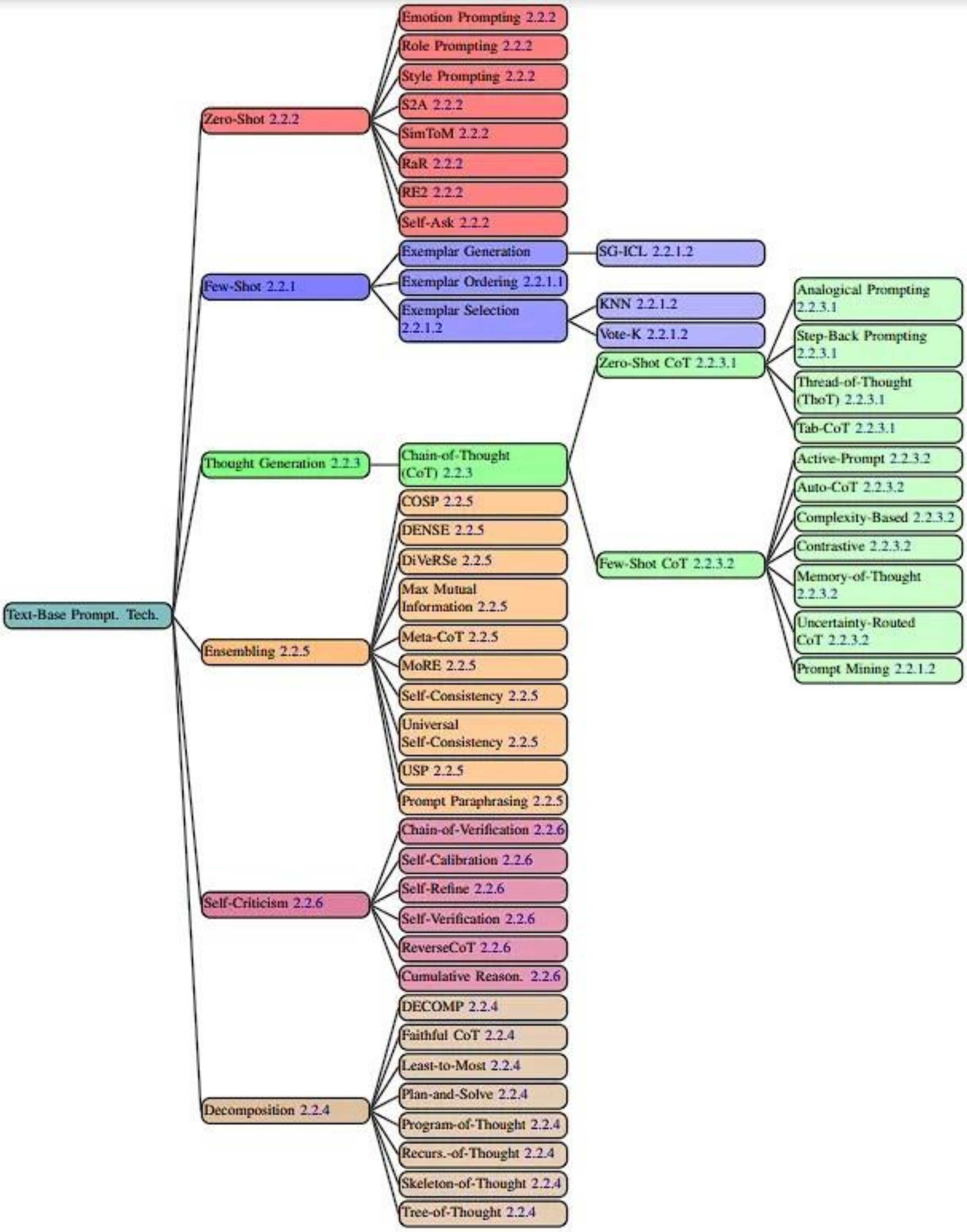
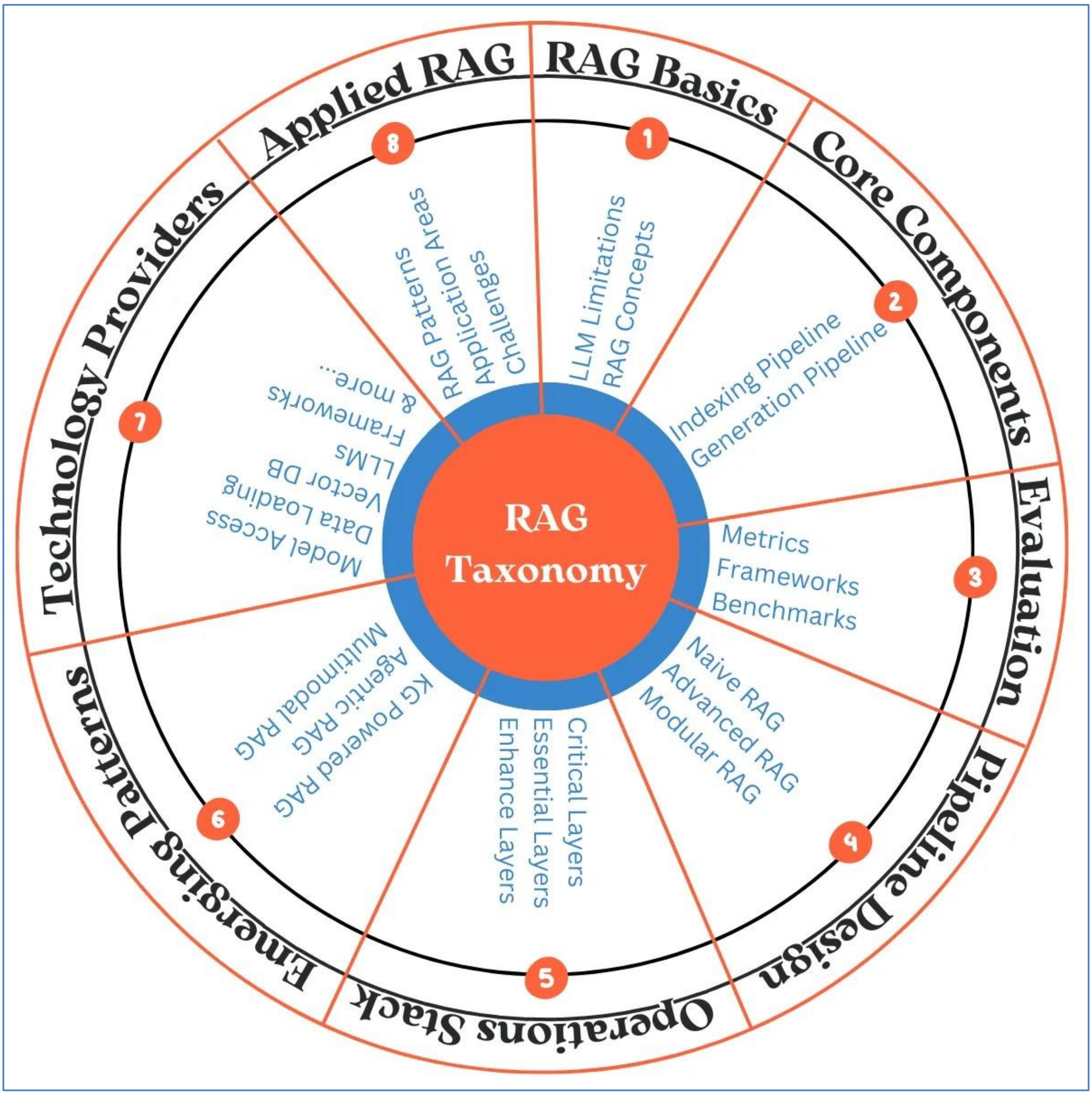


Figure 2.2: All text-based prompting techniques from our dataset.

{Prompt, Text}



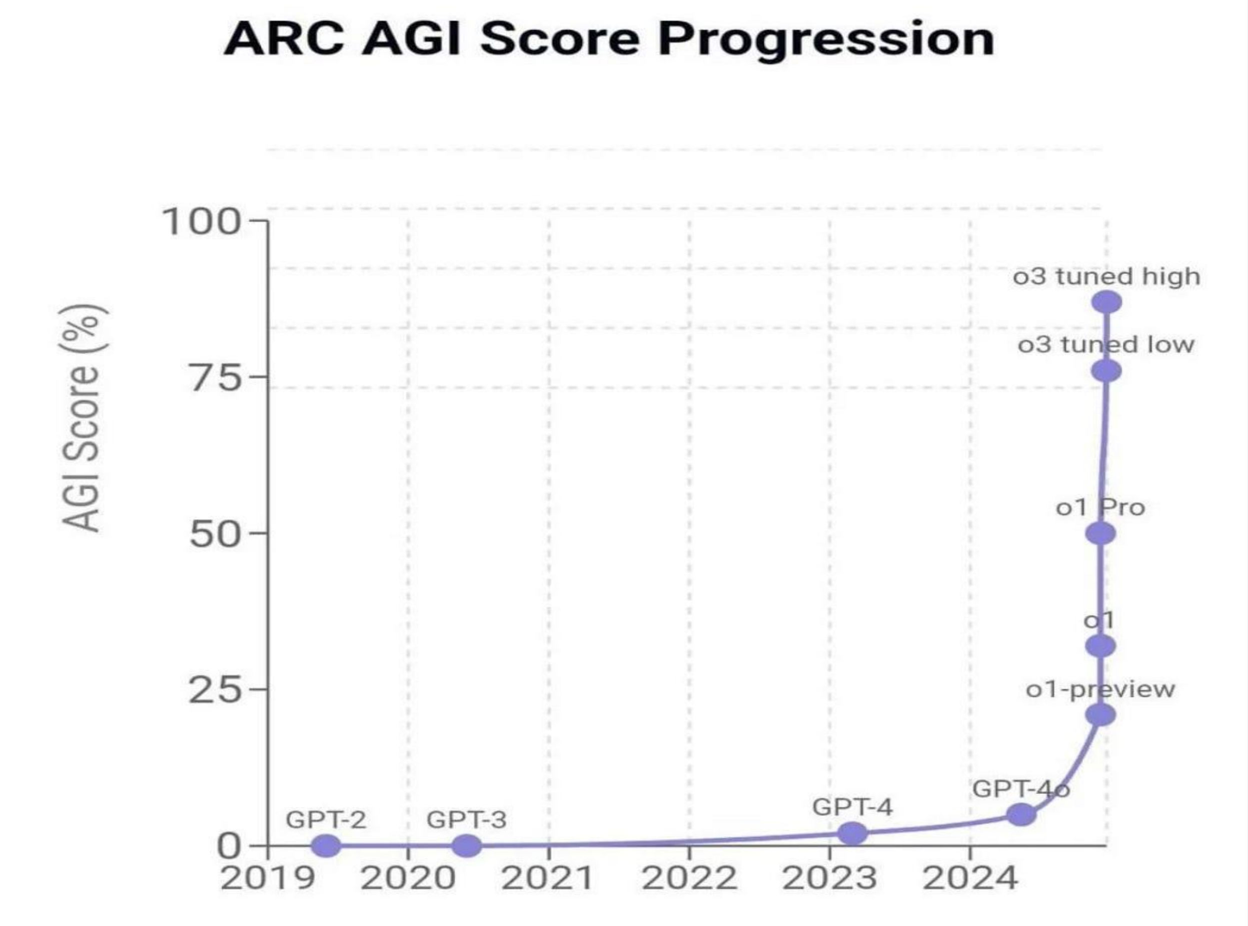
{RAG Taxonomy}

Key findings for our Project

Useful information for SOTA AI models Using

AD. Some more from Medium

Some interesting and useful info from **[Medium (2024)]** – Open public platform https://medium.com



{AGI, Concept, LCM}

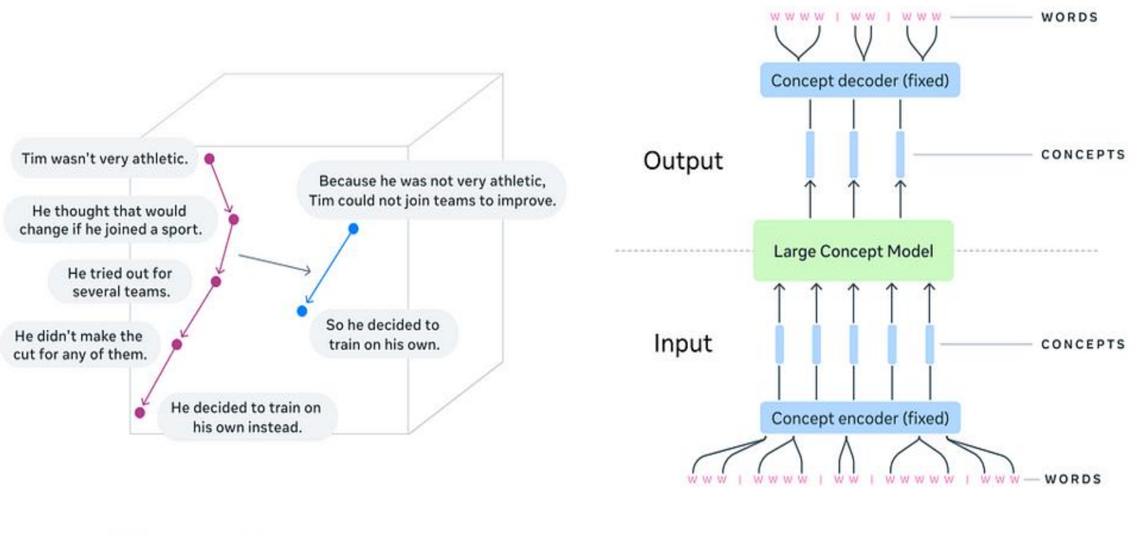
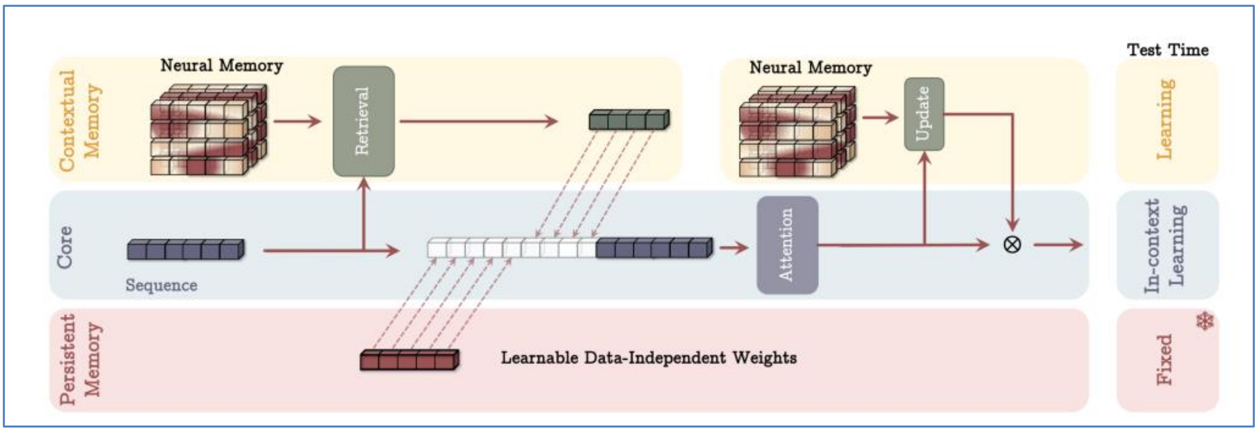

Tim wasn't very athletic.
He thought that would change if he joined a sport.
He tried out for several teams.
He didn't make the cut for any of them.
He decided to train on his own instead.
Because he was not very athletic, Tim could not join teams to improve.
So he decided to train on his own.
WORDS
Concept decoder (fixed)
Output
CONCEPTS
Large Concept Model
Input
CONCEPTS
Concept encoder (fixed)
WORDS

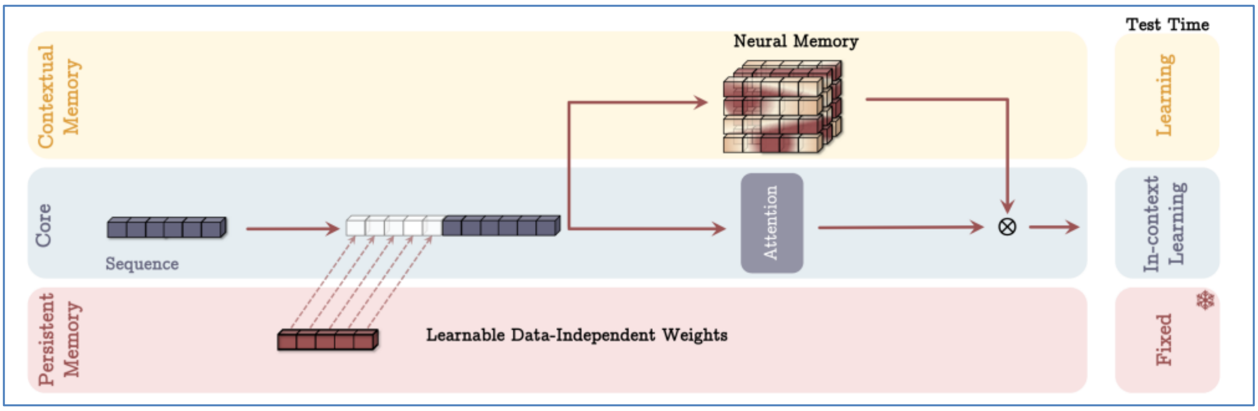
Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an LARGE CONCEPT MODEL (LCM).
⋆: concept encoder and decoder are frozen.

Google Titans - Better alternate for Transformers architecture for LLMs

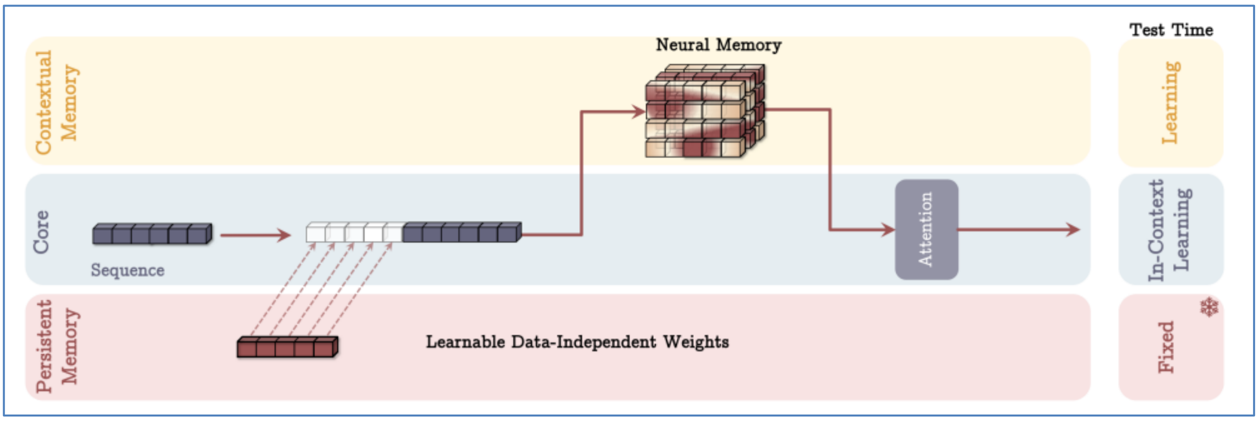
Memory as a Context (MAC):

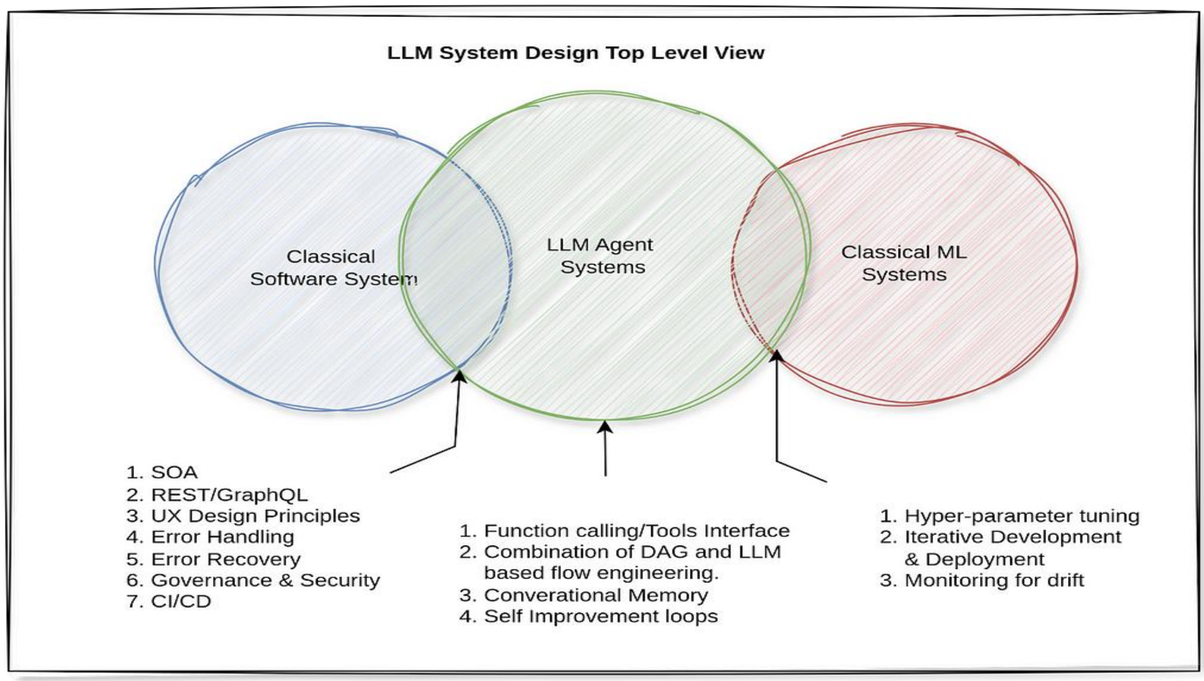


Memory as a Gate (MAG):



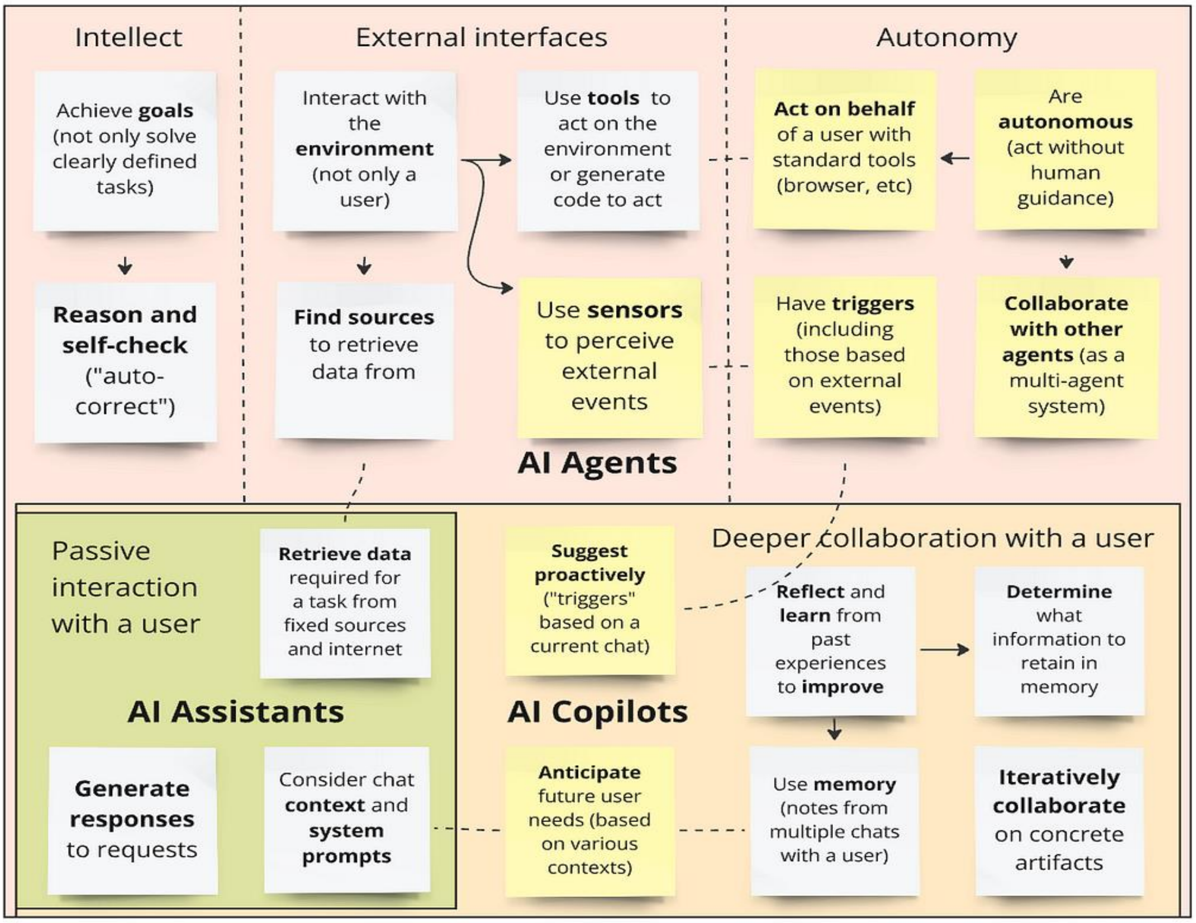
Memory as a Layer (MAL):



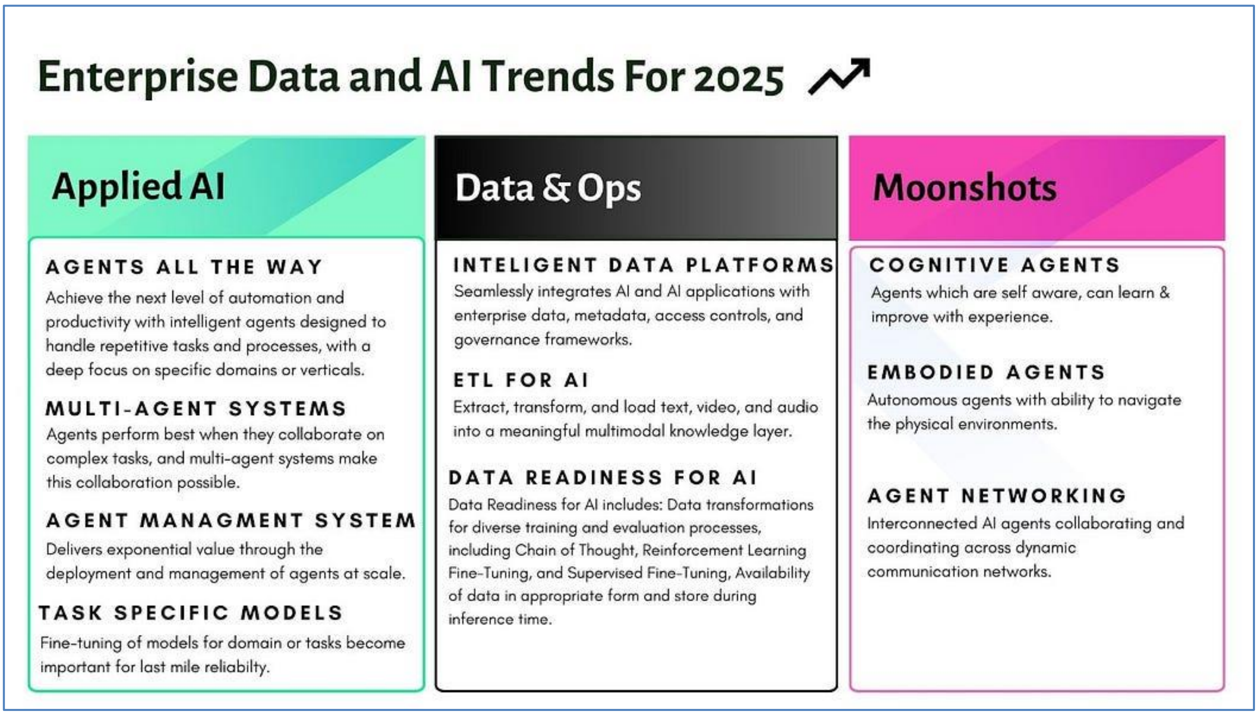

LLM System Design Top Level View
Classical Software System
LLM Agent Systems
Classical ML Systems
1. SOA
2. REST/GraphQL
3. UX Design Principles
4. Error Handling
5. Error Recovery
6. Governance & Security
7. CI/CD
1. Function calling/Tools Interface
2. Combination of DAG and LLM based flow engineering.
3. Converational Memory
4. Self Improvement loops
1. Hyper-parameter tuning
2. Iterative Development & Deployment
3. Monitoring for drift

{LLM Design, Agent, AI, Copilot, GPT}

What AI Agents Can Do in Addition to Capabilities of AI Assistants and Copilots


Intellect
External interfaces
Autonomy
Achieve goals (not only solve clearly defined tasks)
Interact with the environment (not only a user)
Use tools to act on the environment or generate code to act
Act on behalf of a user with standard tools (browser, etc)
Are autonomous (act without human guidance)
Reason and self-check ("auto-correct")
Find sources to retrieve data from
Use sensors to perceive external events
Have triggers (including those based on external events)
Collaborate with other agents (as a multi-agent system)
AI Agents
Passive interaction with a user
Retrieve data required for a task from fixed sources and internet
Suggest proactively ("triggers" based on a current chat)
Deeper collaboration with a user
Reflect and learn from past experiences to improve
Determine what information to retain in memory
AI Assistants
AI Copilots
Generate responses to requests
Consider chat context and system prompts
Anticipate future user needs (based on various contexts)
Use memory (notes from multiple chats with a user)
Iteratively collaborate on concrete artifacts

Capabilities in yellow cards are **not** available in ChatGPT (GPTs) and other widely used AI tools yet



Enterprise Data and AI Trends For 2025

Applied AI

AGENTS ALL THE WAY
Achieve the next level of automation and productivity with intelligent agents designed to handle repetitive tasks and processes, with a deep focus on specific domains or verticals.

MULTI-AGENT SYSTEMS
Agents perform best when they collaborate on complex tasks, and multi-agent systems make this collaboration possible.

AGENT MANAGMENT SYSTEM
Delivers exponential value through the deployment and management of agents at scale.

TASK SPECIFIC MODELS
Fine-tuning of models for domain or tasks become important for last mile reliabilty.

Data & Ops

INTELIGENT DATA PLATFORMS
Seamlessly integrates AI and AI applications with enterprise data, metadata, access controls, and governance frameworks.

ETL FOR AI
Extract, transform, and load text, video, and audio into a meaningful multimodal knowledge layer.

DATA READINESS FOR AI
Data Readiness for AI includes: Data transformations for diverse training and evaluation processes, including Chain of Thought, Reinforcement Learning Fine-Tuning, and Supervised Fine-Tuning, Availability of data in appropriate form and store during inference time.

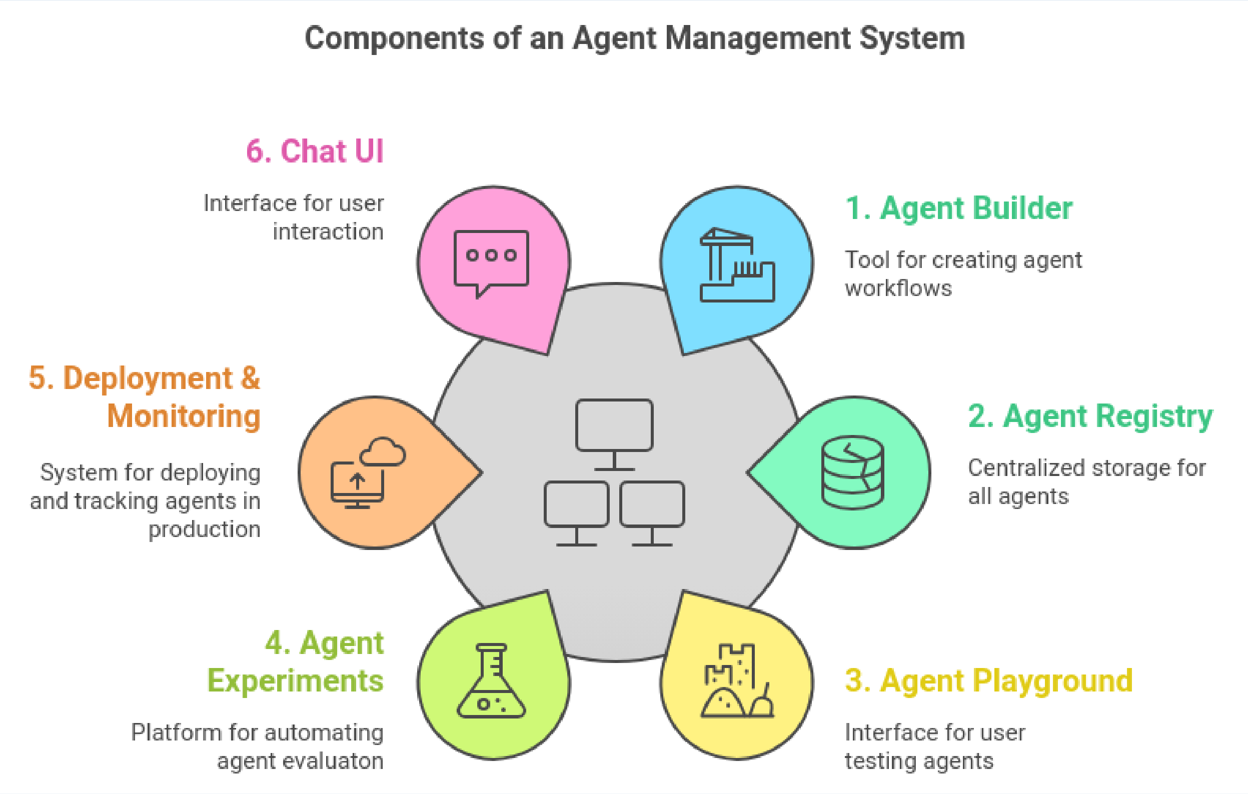
Moonshots

COGNITIVE AGENTS
Agents which are self aware, can learn & improve with experience.

EMBODIED AGENTS
Autonomous agents with ability to navigate the physical environments.

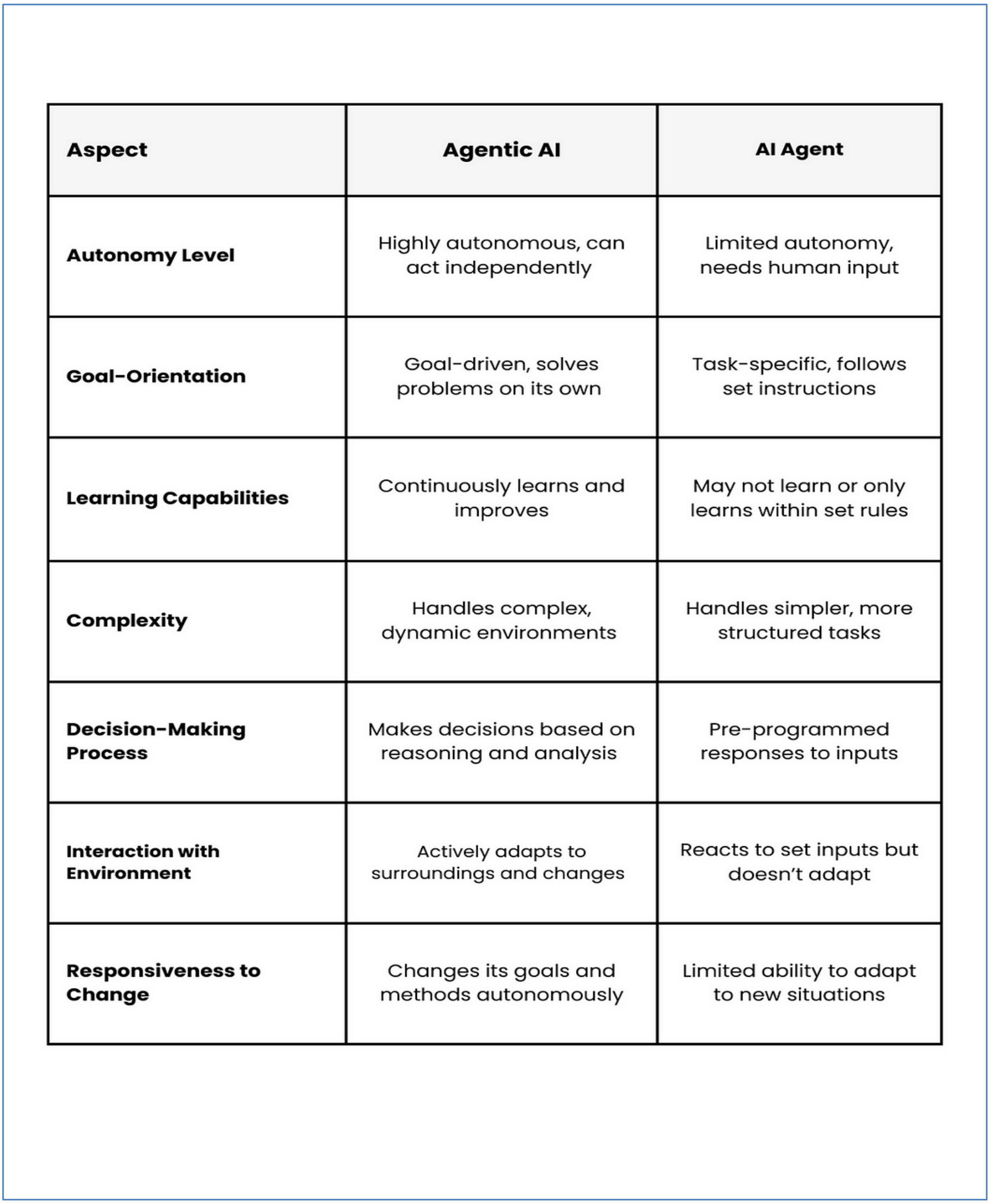
AGENT NETWORKING
Interconnected AI agents collaborating and coordinating across dynamic communication networks.

{Data, AI Trends, Agent, Management System}



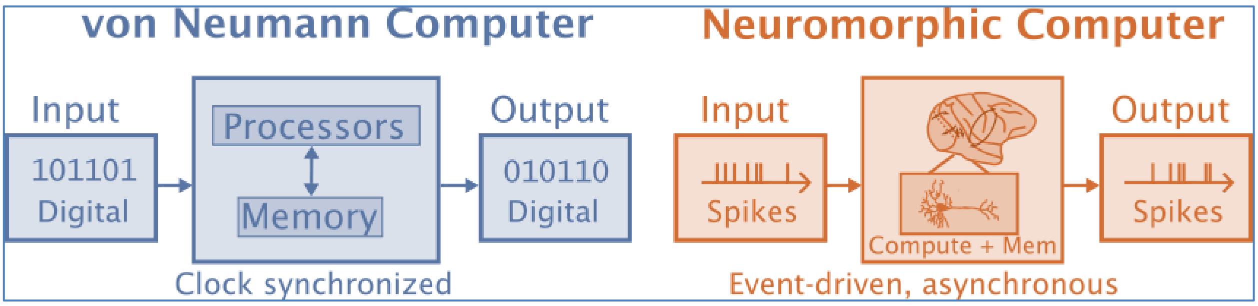
Components of an Agent Management System

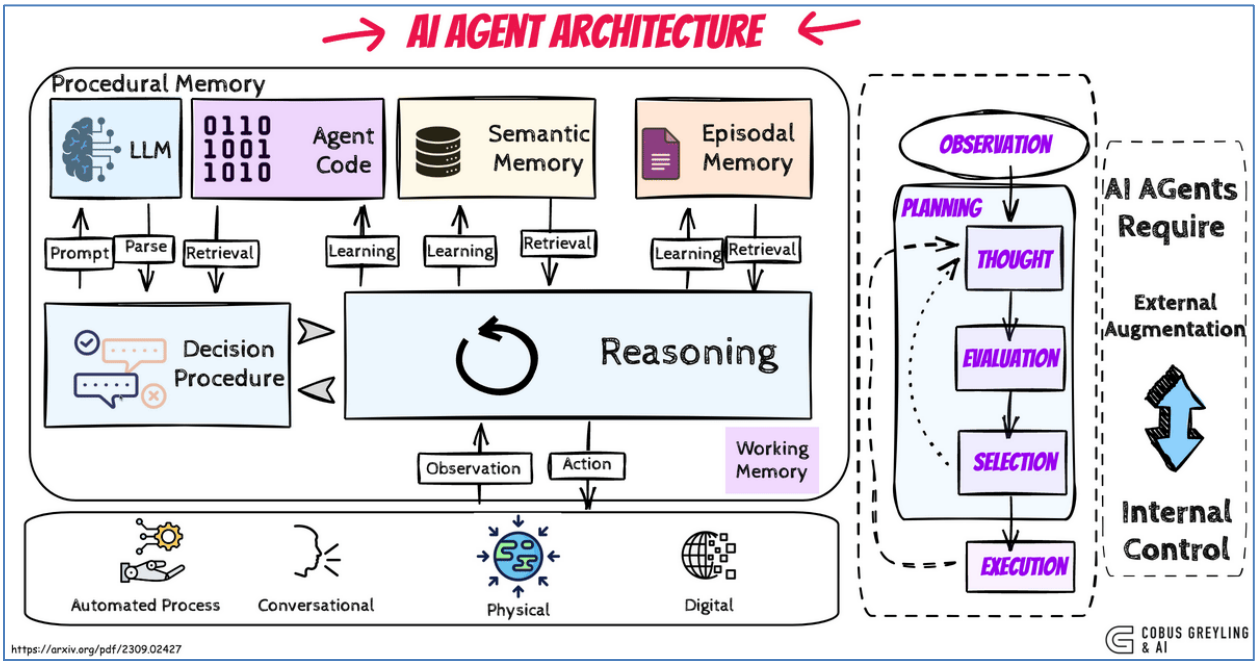
6. Chat UI
Interface for user interaction
1. Agent Builder
Tool for creating agent workflows
5. Deployment & Monitoring
System for deploying and tracking agents in production
2. Agent Registry
Centralized storage for all agents
4. Agent Experiments
Platform for automating agent evaluaton
3. Agent Playground
Interface for user testing agents



Aspect	Agentic AI	AI Agent
Autonomy Level	Highly autonomous, can act independently	Limited autonomy, needs human input
Goal-Orientation	Goal-driven, solves problems on its own	Task-specific, follows set instructions
Learning Capabilities	Continuously learns and improves	May not learn or only learns within set rules
Complexity	Handles complex, dynamic environments	Handles simpler, more structured tasks
Decision-Making Process	Makes decisions based on reasoning and analysis	Pre-programmed responses to inputs
Interaction with Environment	Actively adapts to surroundings and changes	Reacts to set inputs but doesn't adapt
Responsiveness to Change	Changes its goals and methods autonomously	Limited ability to adapt to new situations

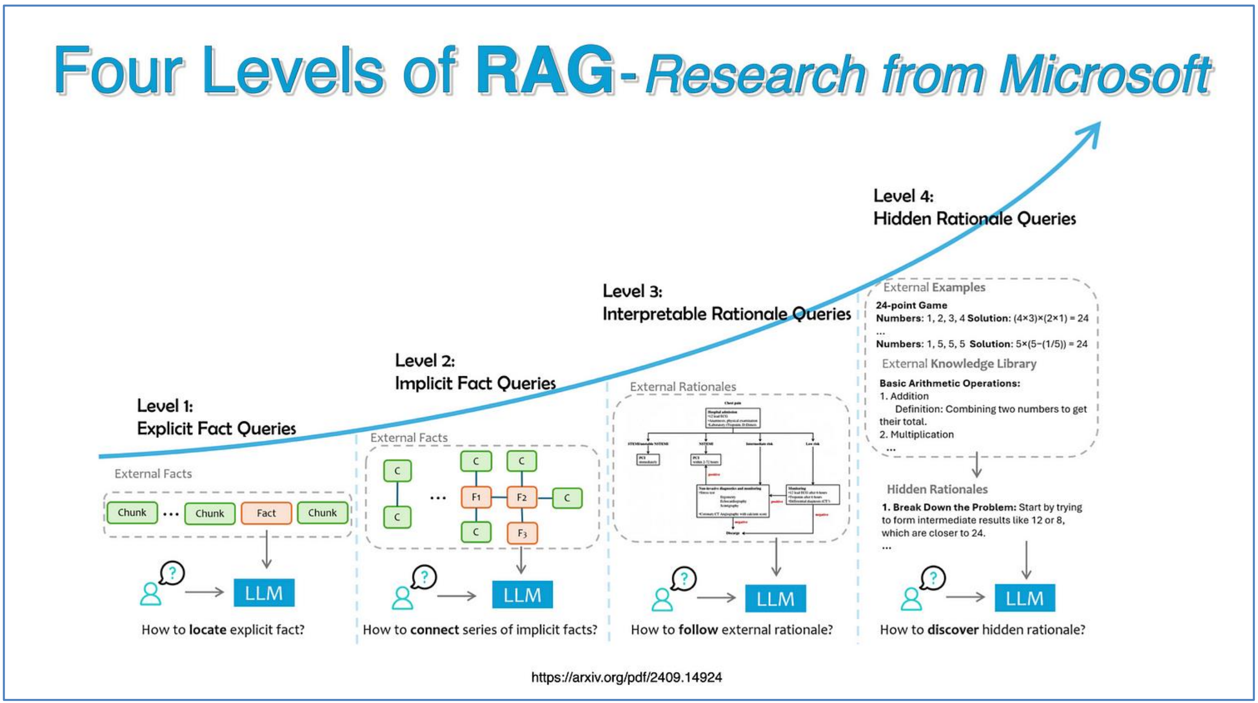
{AI, Agent, Computer}





https://arxiv.org/pdf/2309.02427

{AI Agent Architecture, RAG, Microsoft}

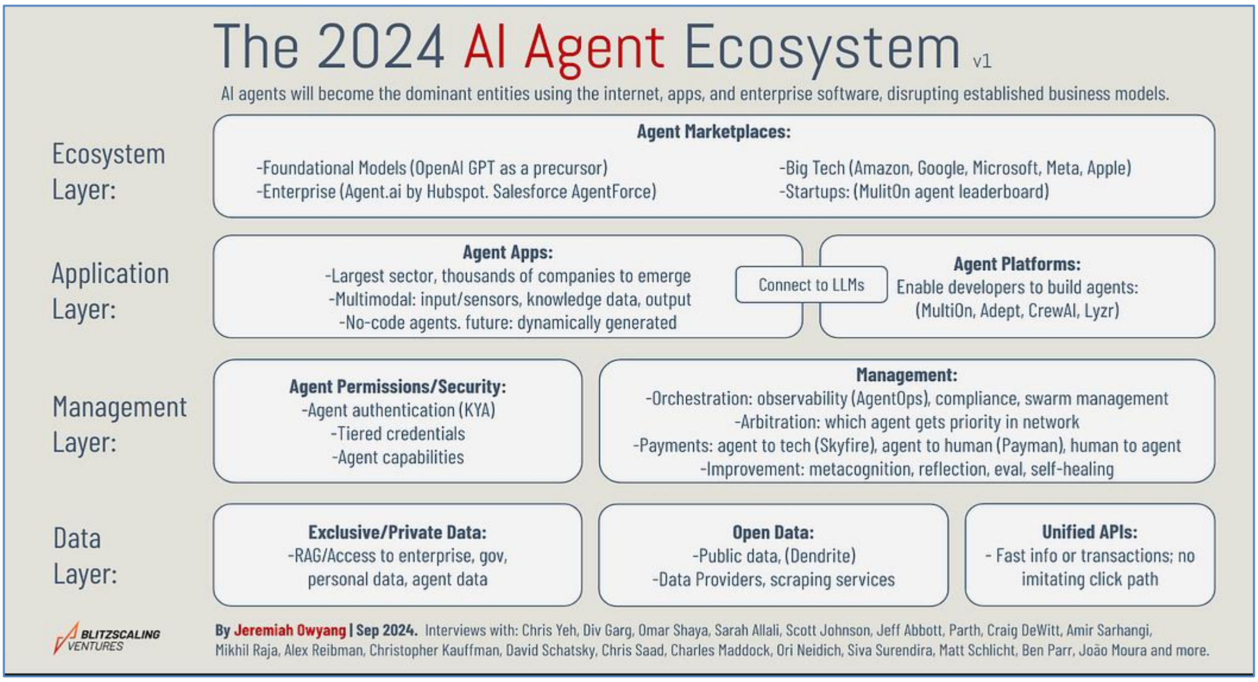

Four Levels of RAG- Research from Microsoft
Level 1:
Explicit Fact Queries
External Facts
Chunk
Chunk
Fact
Chunk
LLM
How to locate explicit fact?
Level 2:
Implicit Fact Queries
External Facts
C
F1
F2
F3
LLM
How to connect series of implicit facts?
Level 3:
Interpretable Rationale Queries
External Rationales
LLM
How to follow external rationale?
Level 4:
Hidden Rationale Queries
External Examples
External Knowledge Library
Hidden Rationales
LLM
How to discover hidden rationale?

24-point Game
Numbers: 1, 2, 3, 4 **Solution:** (4×3)×(2×1) = 24
...
Numbers: 1, 5, 5, 5 **Solution:** 5×(5−(1/5)) = 24

Basic Arithmetic Operations:
1. Addition
Definition: Combining two numbers to get their total.
2. Multiplication
...

1. Break Down the Problem: Start by trying to form intermediate results like 12 or 8, which are closer to 24.
...

https://arxiv.org/pdf/2409.14924



The 2024 AI Agent Ecosystem v1

AI agents will become the dominant entities using the internet, apps, and enterprise software, disrupting established business models.

Ecosystem Layer:

Agent Marketplaces:
-Foundational Models (OpenAI GPT as a precursor)
-Enterprise (Agent.ai by Hubspot. Salesforce AgentForce)
-Big Tech (Amazon, Google, Microsoft, Meta, Apple)
-Startups: (MulitOn agent leaderboard)

Application Layer:

Agent Apps:
-Largest sector, thousands of companies to emerge
-Multimodal: input/sensors, knowledge data, output
-No-code agents. future: dynamically generated

Connect to LLMs

Agent Platforms:
Enable developers to build agents:
(MultiOn, Adept, CrewAI, Lyzr)

Management Layer:

Agent Permissions/Security:
-Agent authentication (KYA)
-Tiered credentials
-Agent capabilities

Management:
-Orchestration: observability (AgentOps), compliance, swarm management
-Arbitration: which agent gets priority in network
-Payments: agent to tech (Skyfire), agent to human (Payman), human to agent
-Improvement: metacognition, reflection, eval, self-healing

Data Layer:

Exclusive/Private Data:
-RAG/Access to enterprise, gov, personal data, agent data

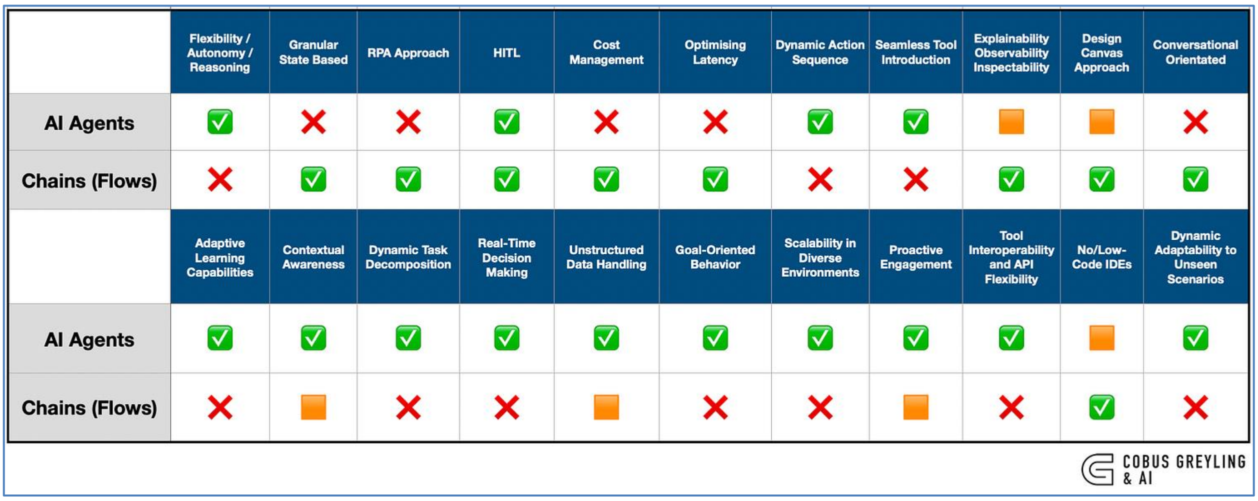
Open Data:
-Public data, (Dendrite)
-Data Providers, scraping services

Unified APIs:
- Fast info or transactions; no imitating click path

BLITZSCALING VENTURES

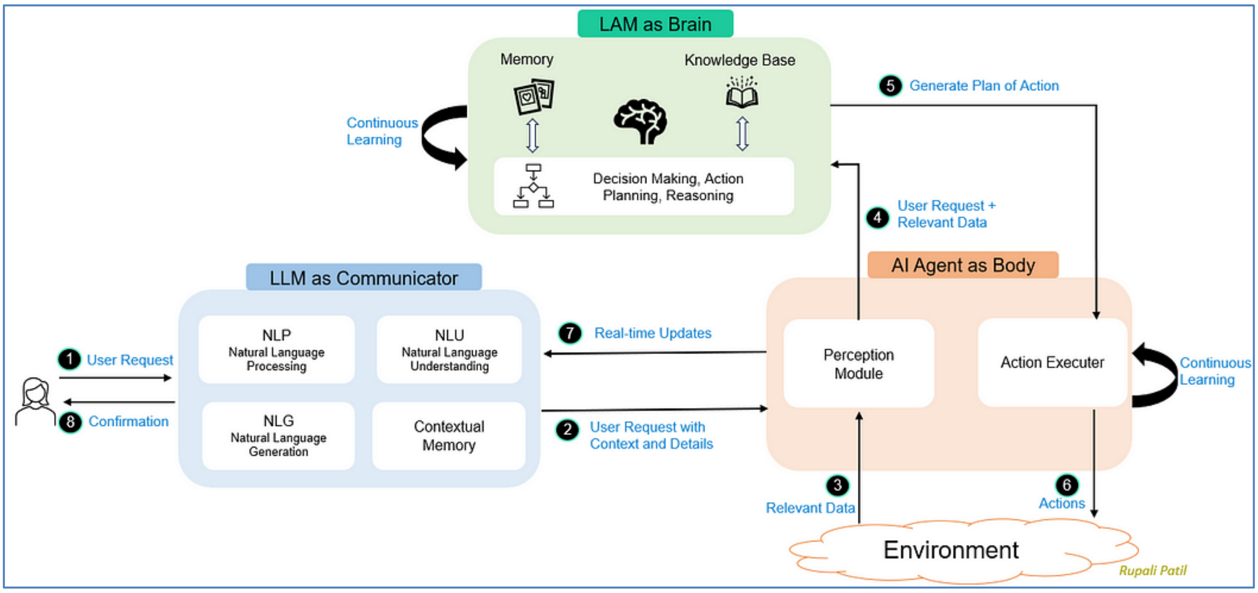
By Jeremiah Owyang | Sep 2024. Interviews with: Chris Yeh, Div Garg, Omar Shaya, Sarah Allali, Scott Johnson, Jeff Abbott, Parth, Craig DeWitt, Amir Sarhangi, Mikhil Raja, Alex Reibman, Christopher Kauffman, David Schatsky, Chris Saad, Charles Maddock, Ori Neidich, Siva Surendira, Matt Schlicht, Ben Parr, João Moura and more.

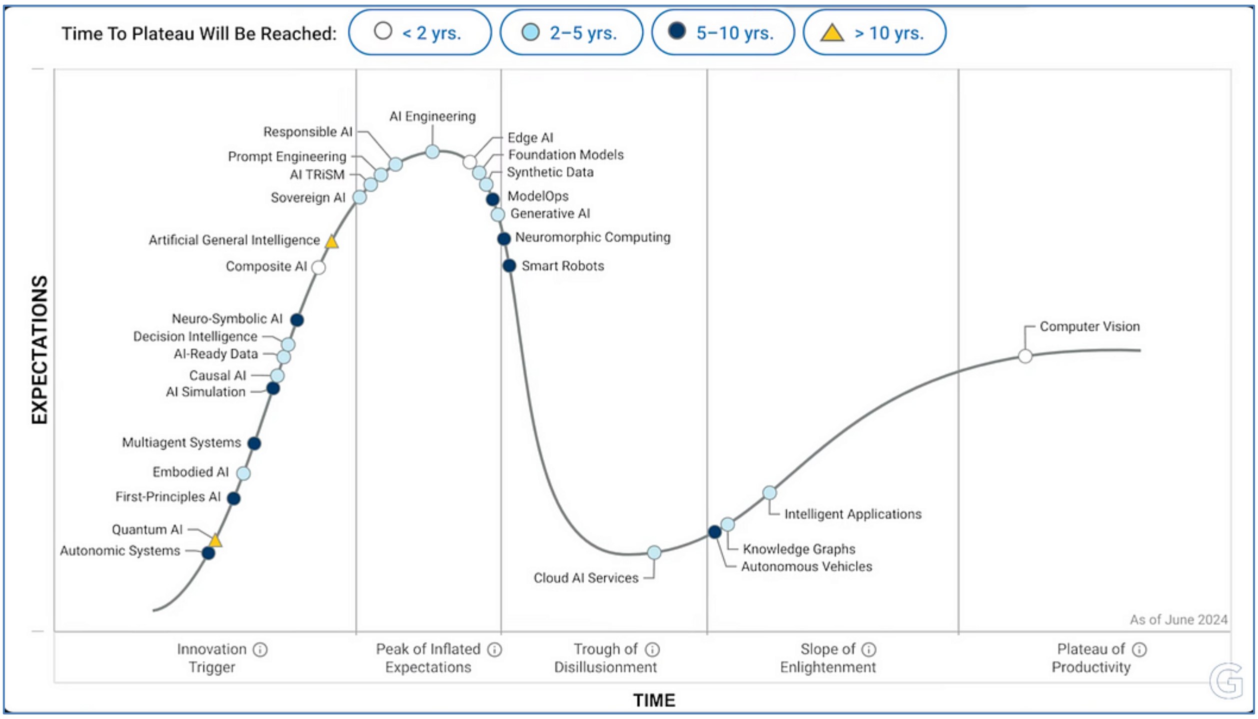
{AI Agent Ecosystem, LLM, LAM}



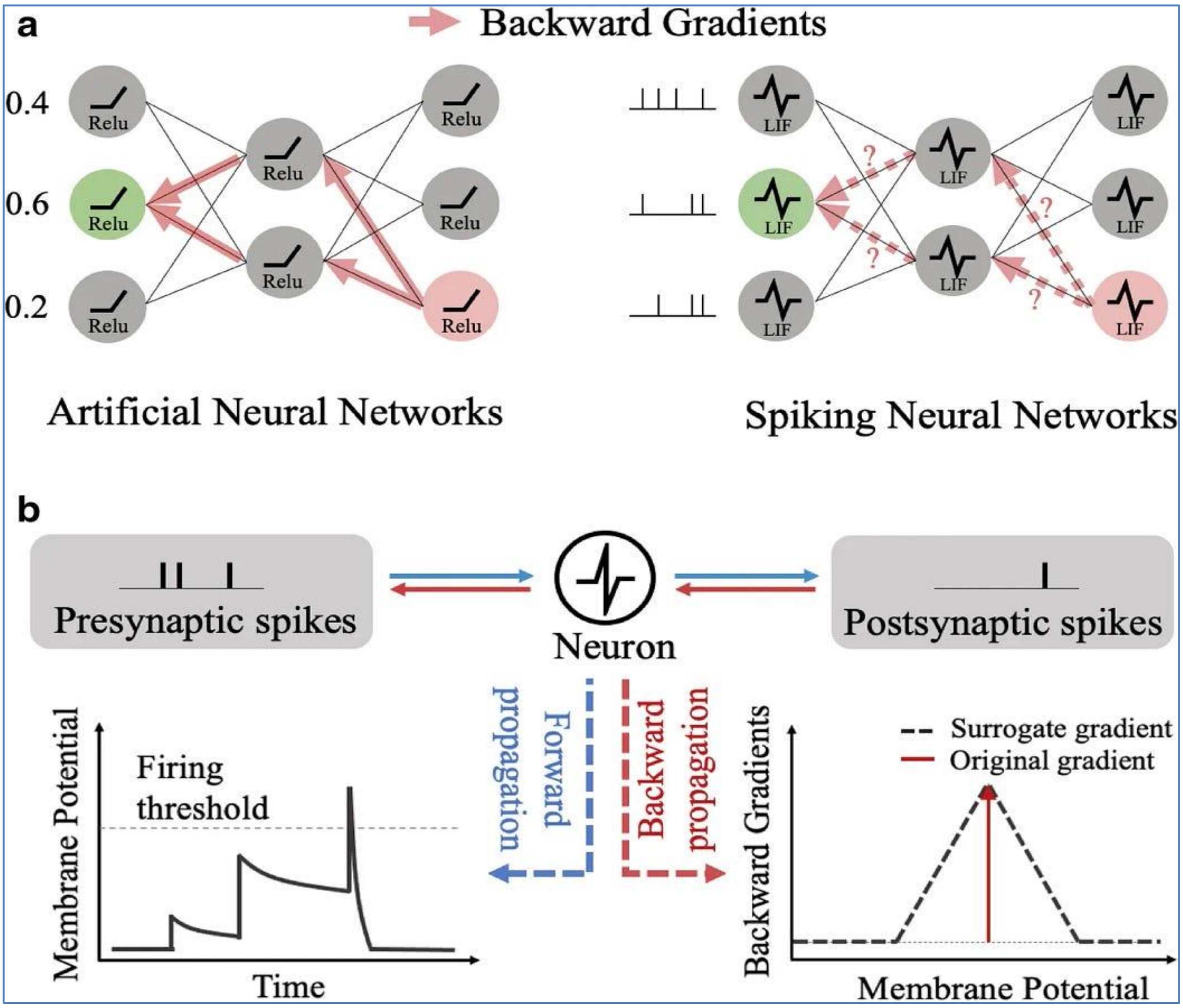
	Flexibility / Autonomy / Reasoning	Granular State Based	RPA Approach	HITL	Cost Management	Optimising Latency	Dynamic Action Sequence	Seamless Tool Introduction	Explainability Observability Inspectability	Design Canvas Approach	Conversational Orientated
AI Agents	✅	❌	❌	✅	❌	❌	✅	✅	🟧	🟧	❌
Chains (Flows)	❌	✅	✅	✅	✅	✅	❌	❌	✅	✅	✅
	Adaptive Learning Capabilities	**Contextual Awareness**	**Dynamic Task Decomposition**	**Real-Time Decision Making**	**Unstructured Data Handling**	**Goal-Oriented Behavior**	**Scalability in Diverse Environments**	**Proactive Engagement**	**Tool Interoperability and API Flexibility**	**No/Low-Code IDEs**	**Dynamic Adaptability to Unseen Scenarios**
AI Agents	✅	✅	✅	✅	✅	✅	✅	✅	✅	🟧	✅
Chains (Flows)	❌	🟧	❌	❌	🟧	❌	❌	🟧	❌	✅	❌

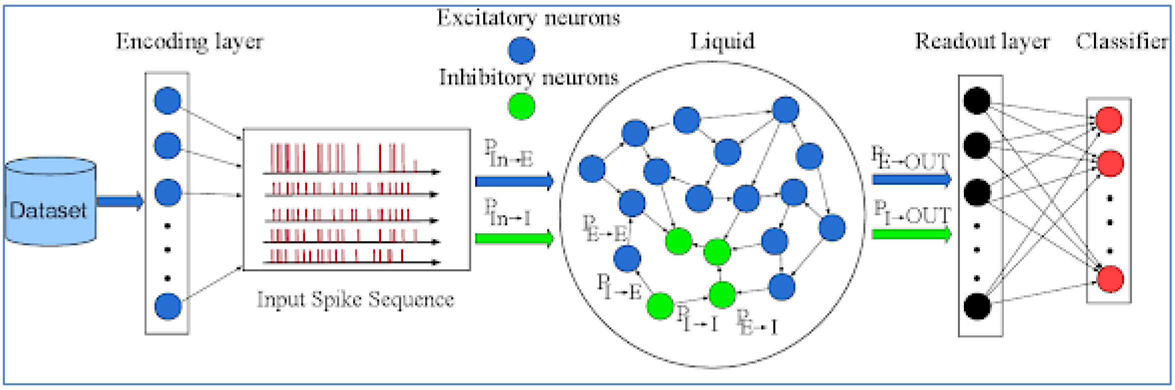
COBUS GREYLING & AI



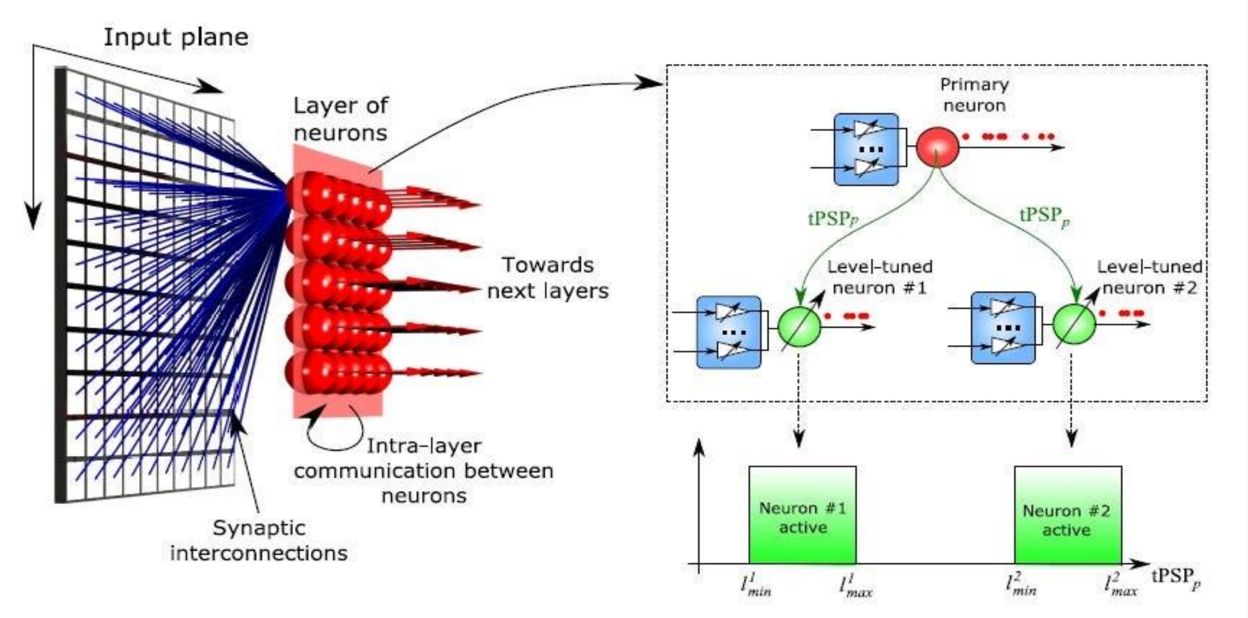


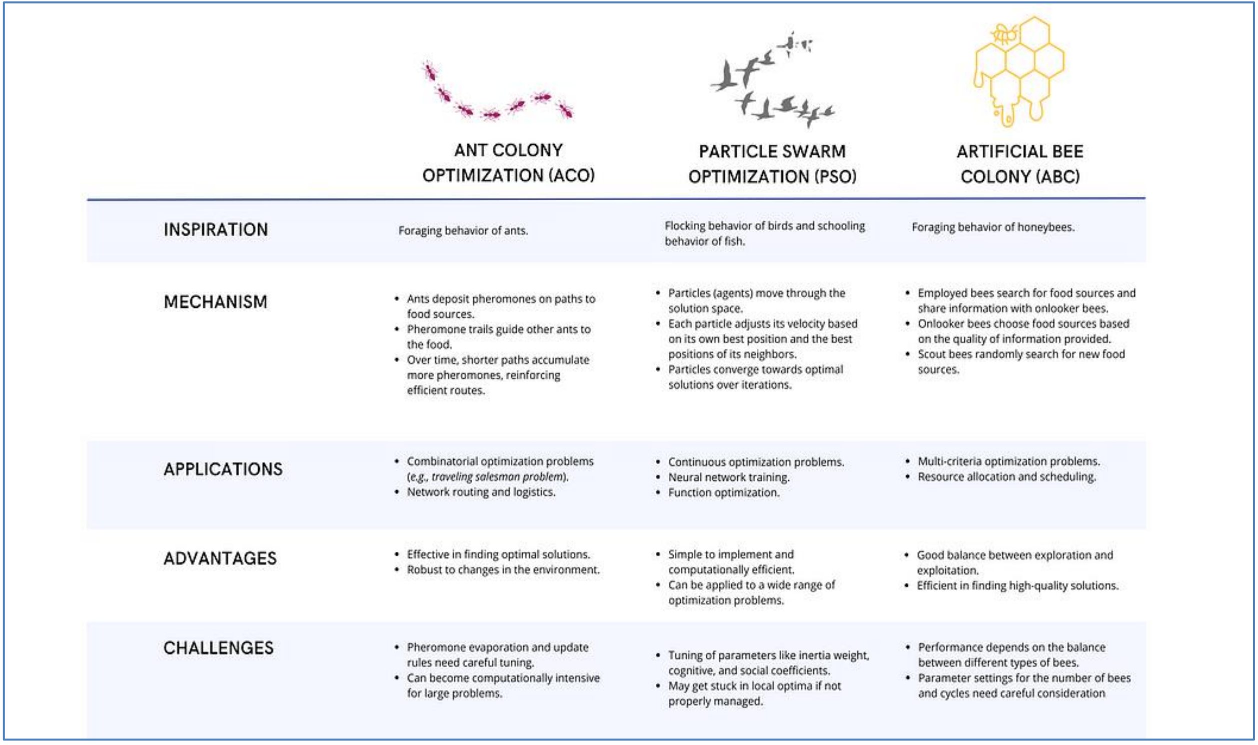
{AI Forecast, Neural network}



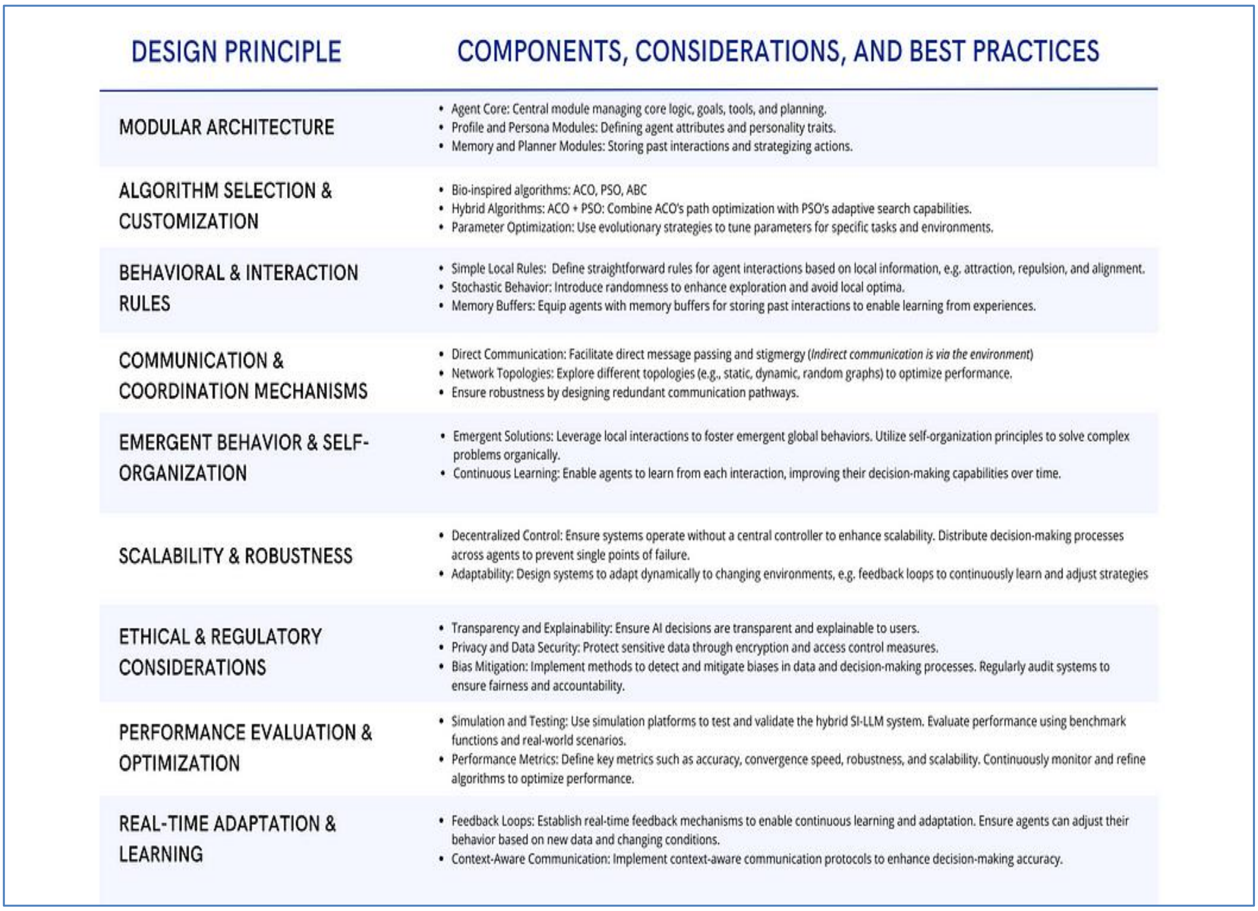


{AI, Neural network, MAS, Swarm}





	ANT COLONY OPTIMIZATION (ACO)	PARTICLE SWARM OPTIMIZATION (PSO)	ARTIFICIAL BEE COLONY (ABC)
INSPIRATION	Foraging behavior of ants.	Flocking behavior of birds and schooling behavior of fish.	Foraging behavior of honeybees.
MECHANISM	• Ants deposit pheromones on paths to food sources. • Pheromone trails guide other ants to the food. • Over time, shorter paths accumulate more pheromones, reinforcing efficient routes.	• Particles (agents) move through the solution space. • Each particle adjusts its velocity based on its own best position and the best positions of its neighbors. • Particles converge towards optimal solutions over iterations.	• Employed bees search for food sources and share information with onlooker bees. • Onlooker bees choose food sources based on the quality of information provided. • Scout bees randomly search for new food sources.
APPLICATIONS	• Combinatorial optimization problems (*e.g., traveling salesman problem*). • Network routing and logistics.	• Continuous optimization problems. • Neural network training. • Function optimization.	• Multi-criteria optimization problems. • Resource allocation and scheduling.
ADVANTAGES	• Effective in finding optimal solutions. • Robust to changes in the environment.	• Simple to implement and computationally efficient. • Can be applied to a wide range of optimization problems.	• Good balance between exploration and exploitation. • Efficient in finding high-quality solutions.
CHALLENGES	• Pheromone evaporation and update rules need careful tuning. • Can become computationally intensive for large problems.	• Tuning of parameters like inertia weight, cognitive, and social coefficients. • May get stuck in local optima if not properly managed.	• Performance depends on the balance between different types of bees. • Parameter settings for the number of bees and cycles need careful consideration



DESIGN PRINCIPLE	COMPONENTS, CONSIDERATIONS, AND BEST PRACTICES
MODULAR ARCHITECTURE	• Agent Core: Central module managing core logic, goals, tools, and planning. • Profile and Persona Modules: Defining agent attributes and personality traits. • Memory and Planner Modules: Storing past interactions and strategizing actions.
ALGORITHM SELECTION & CUSTOMIZATION	• Bio-inspired algorithms: ACO, PSO, ABC • Hybrid Algorithms: ACO + PSO: Combine ACO's path optimization with PSO's adaptive search capabilities. • Parameter Optimization: Use evolutionary strategies to tune parameters for specific tasks and environments.
BEHAVIORAL & INTERACTION RULES	• Simple Local Rules: Define straightforward rules for agent interactions based on local information, e.g. attraction, repulsion, and alignment. • Stochastic Behavior: Introduce randomness to enhance exploration and avoid local optima. • Memory Buffers: Equip agents with memory buffers for storing past interactions to enable learning from experiences.
COMMUNICATION & COORDINATION MECHANISMS	• Direct Communication: Facilitate direct message passing and stigmergy (*Indirect communication is via the environment*) • Network Topologies: Explore different topologies (e.g., static, dynamic, random graphs) to optimize performance. • Ensure robustness by designing redundant communication pathways.
EMERGENT BEHAVIOR & SELF-ORGANIZATION	• Emergent Solutions: Leverage local interactions to foster emergent global behaviors. Utilize self-organization principles to solve complex problems organically. • Continuous Learning: Enable agents to learn from each interaction, improving their decision-making capabilities over time.
SCALABILITY & ROBUSTNESS	• Decentralized Control: Ensure systems operate without a central controller to enhance scalability. Distribute decision-making processes across agents to prevent single points of failure. • Adaptability: Design systems to adapt dynamically to changing environments, e.g. feedback loops to continuously learn and adjust strategies
ETHICAL & REGULATORY CONSIDERATIONS	• Transparency and Explainability: Ensure AI decisions are transparent and explainable to users. • Privacy and Data Security: Protect sensitive data through encryption and access control measures. • Bias Mitigation: Implement methods to detect and mitigate biases in data and decision-making processes. Regularly audit systems to ensure fairness and accountability.
PERFORMANCE EVALUATION & OPTIMIZATION	• Simulation and Testing: Use simulation platforms to test and validate the hybrid SI-LLM system. Evaluate performance using benchmark functions and real-world scenarios. • Performance Metrics: Define key metrics such as accuracy, convergence speed, robustness, and scalability. Continuously monitor and refine algorithms to optimize performance.
REAL-TIME ADAPTATION & LEARNING	• Feedback Loops: Establish real-time feedback mechanisms to enable continuous learning and adaptation. Ensure agents can adjust their behavior based on new data and changing conditions. • Context-Aware Communication: Implement context-aware communication protocols to enhance decision-making accuracy.

{AI Design, MAS}

Key findings for our Project

Useful information about (from) AI R&Ds – MASs, Agency etc.

References

1. **Abbas et al. (2021)** Abbas, A. et al. The power of quantum neural networks. Nat. Comput. Sci. 1, 403 (2021)
2. **Active Inference Institute (2024)** AII (Team) Active Inference Institute & Active Inference Ecosystem (2024, v2). DOI: 10.5281/zenodo.14108992
3. **Ahmad, Subutai & Scheinkman (2019)** Ahmad, Subutai and Scheinkman How Can We Be So Dense? The Robustness of Highly Sparse Representations". In: ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning.
4. **AI Portal (2019)** http://www.aiportal.ru/ (*since 2022 closed*)
5. **AI progress (2022)** AI progress. National Academy Committee on Automation and the U.S. Workforce, July 6th 2022
6. **AI100 (2021)** Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report." Stanford University, Stanford, CA, Sept 2021.
7. **AIIR (2024)** Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, "The AI Index 2024 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.
8. **Akiba et al. (2024)** Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, David Ha. Evolutionary Optimization of Model Merging Recipes. arXiv:2403.13187v1 [cs.NE] 19 Mar 2024
9. **Alabdulmohsin, Tran & Dehghani (2024)** Ibrahim Alabdulmohsin, Vinh Q. Tran, Mostafa Dehghani. Fractal Patterns May Unravel the Intelligence in Next-Token Prediction. arXiv:2402.01825v1 [cs.CL] 2 Feb 2024
10. **Albalak et al. (2024)** Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto and William Yang Wang. A Survey on Data Selection for Language Models. arXiv:2402.16827v2 [cs.CL] 8 Mar 2024
11. **Albarracin et al. (2022)** Mahault Albarracin, Daphne Demekas, Maxwell J.D. Ramstead, and Conor Heins. "Epistemic communities under active inference". In: Entropy 24.4 (2022), p. 476.
12. **Albarracin et al. (2024)** Mahault Albarracin, Riddhi J. Pitliya, Toby St. Clere Smithe, Daniel Ari Friedman, Karl Friston, Maxwell J. D. Ramstead. Shared Protentions in Multi-Agent Active Inference. Entropy 2024, 26, 303.
13. **Anderson & Lebiere (2003)** Anderson, J. R. & Lebiere, C. The newell test for a theory of cognition. Behavioral and brain Sciences 26, 587–601 (2003).
14. **Anderson (1983)** Anderson, J. The Architecture of Cognition (1983).
15. **Anderson (1990)** Anderson, J. R. The adaptive character of thought (Erlbaum, Hillsdale, NJ, 1990).
16. **Andersson et al. (2009)** Jesper Andersson, Rogério de Lemos, Sam Malek, and Danny Weyns. Modeling Dimensions of Self-Adaptive Software Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 27–47.
17. **Andreessen (2023)** Marc Andreessen. Why AI Will Save the World. Andreessen Horowitz (a16z). https://a16z.com/2023/06/06/ai-will-save-the-world/
18. **Anthropic (2022)** Anthropic. Constitutional AI: Harmlessness from AI Feedback https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback
19. **Anthropic (2023a)** Anthropic. Claude-2. https://www.anthropic.com/index/claude-2

20. **Anthropic (2023b)** Anthropic. Model Card and Evaluations for Claude Models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf
21. **Anthropic (2024)** Anthropic. Mapping the Mind of a Large Language Model. https://www.anthropic.com/news/mapping-mind-language-model. 21 May 2024
22. **Arcas & Norvig (2023)** Blaise Agüera Y Arcas & Peter Norvig. Artificial General Intelligence Is Already Here. https://www.noemamag.com/artificial-general-intelligence-is-already-here/ October 10, 2023
23. **Ardelt (2004)** Ardelt, M. Wisdom as expert knowledge system: a critical review of a contemporary operationalization of an ancient concept. Human Development, 47, 257–287. (2004)
24. **Arslan (2024)** Suayb S. Arslan. Artificial Human Intelligence: The role of Humans in the Development of Next Generation AI arXiv:2409.16001v1 [cs.AI] 24 Sep 2024
25. **Aru et al. (2020)** Aru, J., Suzuki, M. & Larkum, M. E. Cellular mechanisms of conscious processing. Trends Cogn. Sci. 24, 814–825 (2020)
26. **Ashby (1956)** W. Ross Ashby. An Introduction to Cybernetics. London: Chapman & Hall, 1956.
27. **ASI Alliance (2024)** Artificial Superintelligence (ASI) Alliance Vision Paper**.** Building Decentralized Artificial Superintelligence**.** SingularityNET, Fetch.ai, Ocean Protocol - Apr 2024
28. **Assran et al. (2023)** Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, Nicolas Ballas, Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. arXiv:2301.08243v3 [cs.CV] 13 Apr 2023
29. **ATI (2022)** The Alan Turing Institute https://www.turing.ac.uk
30. **BAAI (2023)** BAAI (Beijing Academy of Artificial Intelligence). Wudao Aquila Large Language Model. https://github.com/FlagAI-Open/FlagAI/blob/master/examples/Aquila/README_en.md. (June 2023)
31. **Baars (1988)** B. J. Baars, A Cognitive Theory of Consciousness (Cambridge University Press, Cambridge, United Kingdom, 1988).
32. **Baars (1994)** Bernard J Baars. A global workspace theory of conscious experience. Consciousness in philosophy and cognitive neuroscience, pp. 149–171, 1994.
33. **Baars (1997)** B. J. Baars, In the Theater of Consciousness (Oxford University Press, New York, NY, 1997)
34. **Baars (2005)** Bernard J Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research, 150:45–53, 2005.
35. **BaGuaLu (2022)** BaGuaLu: Targeting Brain Scale Pretrained Models with over 37 Million Cores. PPoPP ’22, April 2–6, 2022, Seoul, Republic of Korea
36. **Baidu (2023)** Baidu Showcases Major AI Developments at WAVE SUMMIT 2023: ERNIE Bot Plugins, PaddlePaddle V2.5, and AI Coding Assistant. https://www.prnewswire.com/news-releases/baidu-showcases-major-ai-developments-at-wave-summit-2023-ernie-bot-plugins-paddlepaddle-v2-5--and-ai-coding-assistant-301903604.html (2023)
37. **Baltes & Smith (2008)** Baltes, P. B. and Smith, J. The fascination of wisdom: Its nature, ontogeny, and function. Perspectives on Psychological Science, 3, 56–64. (2008)
38. **Baraba´si & Baraba´si, (2020)** Da´niel L. Baraba´si, Albert-La´szlo´ Baraba´si , A Genetic Model of the Connectome, Neuron (2019)
39. **Baraba’si (2016)** A.-L. Baraba’si, Network Science (Cambridge University Press, 2016).
40. **Bardes et al. (2021)** Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In International Conference on Learning Representations (ICLR 2022). arXiv preprint arXiv:2105.04906.
41. **Barnes & Hutson (2024)** Emily Barnes and James Hutson. AI and the Cognitive Sense of Self. Journal of Intelligent Communication | Volume 3 | Issue 1

42. **Barrat et al. (2008)** A. Barrat, M. Barthelemy, and A. Vespignani, Dynamical processes on complex networks (Cambridge University Press, 2008).
43. **Barreto et al. (2019)** Barreto, A., Borsa, D., Hou, S., Comanici, G., Aygün, E., Hamel, P., Toyama, D., Hunt, J., Mourad, S., Silver, D., Precup D. The option keyboard: Combining skills in reinforcement learning. In: Proceedings of the Conference on Neural Information Processing Systems. 2019
44. **Barrett (2017)** Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci. 12, 1833 (2017)
45. **Barrett et al. (2023)** Lisa Feldman Barrett, Christiana Westlin , Jordan E. Theriault, Yuta Katsumi, Alfonso Nieto-Castanon, Aaron Kucyi, Sebastian F. Ruf, Sarah M. Brown, Misha Pavel, Deniz Erdogmus, Dana H. Brooks, Karen S. Quigley, Susan Whitfield-Gabrieli. Improving the study of brain-behavior relationships by revisiting basic assumptions. Trends in Cognitive Sciences, OPINION| VOLUME 27, ISSUE 3, P246-257, MARCH 2023, Published:February 02, 2023
46. **Barrow & Tipler (1986)** J.D. Barrow and F.J. Tipler. The Anthropic Cosmological Principle. Oxford University Press, 1986.
47. **Barth´elemy (2011)** M. Barth´elemy, Spatial networks, Physics Reports 499, 1 (2011).
48. **Basseches (1980)** Basseches, M. Dialectical schemata: A framework for the empirical study of the development of dialectical thinking. Human Development, 23, 400-421. (1980)
49. **Bekenstein (1981)** Jacob D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. Phys. Rev. D, 23(2):287-298, Jan 1981.
50. **Benaich & ASC (2023)** Nathan Benaich & Air Street Capital. State of AI Report 2023. October 12, 2023. stateof.ai 2023
51. **Benaich & Hogarth (2022)** Nathan Benaich and Ian Hogarth. State of AI Report October 11, 2022. stateof.ai 2022
52. **Bengio et al. (2025)** Y. Bengio et al. (90+ authors) “International AI Safety Report” (DSIT 2025/001)
53. **Bennett (2023)** Bennett, M.T.: Emergent causality and the foundation of consciousness. In: Hammer, P., Alirezaie, M., Stranneg˚ard, C. (eds.) Artificial General Intelligence, pp. 52–61. Springer, Cham (2023)
54. **Bennett (2024a)** Bennett, M.T.: Computational dualism and objective superintelligence. In: Th´orisson, K.R., Isaev, P., Sheikhlar, A. (eds.) Artificial General Intelligence, pp. 22–32. Springer, Cham (2024)
55. **Bennett (2024b)** Bennett, M.T.: Is complexity an illusion? In: Th´orisson, K.R., Isaev, P., Sheikhlar, A. (eds.) Artificial General Intelligence, pp. 11–21. Springer, Cham (2024)
56. **Bennett (2024c)** Bennett, M.T.: Multiscale Causal Learning. Manuscript under review (2024)
57. **Bennett, Welsh & Ciaunica (2024)** Michael Timothy Bennett, Sean Welsh, Anna Ciaunica. Why Is Anything Conscious? arXiv:2409.14545v2 [cs.AI] 2 Nov 2024
58. **Bereska & Gavves (2024)** Leonard Bereska and Efstratios Gavves. “Mechanistic Interpretability for AI Safety – A Review”. In: (Apr. 2024).
59. **Berwick & Chomsky (2016)** Robert C Berwick and Noam Chomsky. Why Only Us: Language and Evolution. MIT press, 2016.
60. **Bettencourt at al. (2007)** Luis M. A. Bettencourt, José Lobo, Dirk Helbing, Christian Kuhnert, and Geoffrey B. West. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences, 104(17):7301-7306, April 2007.
61. **Betzel & Bassett (2017)** Betzel, R.F., and Bassett, D.S. Generative models for network neuroscience: prospects and promise. J. R. Soc. Interface 14, 20170623.
62. **Bhaskar & Kuppan (2024)** Deepak Bhaskar Acharya, Karthigeyan Kuppan. Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive Survey. IEEE Access. VOLUME 4, 2016 (2024)

63. **Bhoopchand et al. (2023)** Avishkar Bhoopchand, Bethanie Brownfield, Adrian Collister, Agustin Dal Lago, Ashley Edwards, Richard Everett, Alexandre Fréchette, Yanko Gitahy Oliveira, Edward Hughes, Kory W. Mathewson, Piermaria Mendolicchio, Julia Pawar, Miruna Pîslar, Alex Platonov, Evan Senter, Sukhdeep Singh, Alexander Zacherl, Lei M. Zhang. Learning few-shot imitation as cultural transmission. Nature Communications | (2023) 14:7536
64. **Bianconi (2018)** G. Bianconi, Multilayer networks: structure and function (Oxford University Press, 2018).
65. **Biedma et al. (2024)** Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, Xing Xie. Beyond Human Norms: Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches. arXiv:2404.12744v1 [cs.CL] 19 Apr 2024
66. **Binz & Schulz (2024)** Binz, M. & Schulz, E. Turning large language models into cognitive models (2024).
67. **Binz et al. (2023)** Binz, M. et al. Meta-learned models of cognition. Behavioral and Brain Sciences 1–38 (2023).
68. **Binz et al. (2024)** Marcel Binz et al. (50+ authors) Centaur: a foundation model of human cognition. arXiv:2410.20268v2 18 Nov 2024
69. **Birch (2022a)** Birch, J. Materialism and the moral status of animals. The Philosophical Quarterly, 72(4), 2022, pp.795–815.
70. **Birch (2022b)** Birch, J. The search for invertebrate consciousness. Noˆus, 56, 2022. pp.133–153
71. **Birch et al. (2020)** Birch, J., Schnell, A. K., & Clayton, N. S., 2020. Dimensions of animal consciousness. Trends in Cognitive Sciences, 24(10), 2020, pp.789–801.
72. **Blake et al. (2019)** Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, and Surya Ganguli. "A deep learning framework for neuroscience". In: Nat. Neurosci. 22.11 (2019), pp. 1761–1770.
73. **Block (1995)** Block, N. On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18, 1995, pp.227–247.
74. **Block (1996)** Block, N. Mental paint and mental latex. Philosophical Issues, 7, 1996, pp.19–49.
75. **Block (2002)** Block, N. Some concepts of consciousness. Philosophy of Mind: Classical and Contemporary Readings. 2002, pp.206–218.
76. **Block (2007)** Block, N. Consciousness, accessibility, and the mesh between psychology and neuroscience. Behavioral and Brain Sciences, 30, 2007, pp.481–499.
77. **Block (2011)** Block N. Perceptual consciousness overflows cognitive access. Trends Cogn Sci. 15:567–575.
78. **Block (2023)** Block, N. The Border Between Seeing and Thinking. Oxford University Press. 2023
79. **Blum & Blum (2021)** M. Blum, L. Blum, A theoretical computer science perspective on consciousness. JAIC 8, 1–42 (2021).
80. **Blum & Blum (2022)** Lenore Blum and Manuel Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. PNAS 2022 Vol. 119 No. 21
81. **Bolloba's (2001)** Bolloba's, B. Random Graphs (Cambridge University Press, 2001).
82. **Bostrom & Sandberg (2008)** Nick Bostrom and Anders Sandberg. Whole Brain Emulation: A Roadmap. en. Tech. rep. Future of Humanity Institute, 2008.
83. **Bostrom (1998)** Nick Bostrom. Singularity and predictability. http://hanson.gmu.edu/vc.html#bostrom, 1998.
84. **Bostrom (2002)** N Bostrom. Existential Risks: analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology, 9, 2002.

85. **Bostrom (2014)** Bostrom, N. Superintelligence: Paths, Dangers, Strategies. Oxford University Press., 2014
86. **Brinkmann et al. (2023)** Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, Iyad Rahwan. Machine Culture. Nature Human Behaviour 7, 1855–1868 (2023)
87. **Brown et al. (2019)** Brown, R., Lau, H., & LeDoux, J. E. Understanding the higher-order approach to consciousness. Trends in cognitive sciences, 23, 2019, pp.754–768.
88. **Brown et al. (2020)** Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
89. **Brynjolfsson & McAfee (2014)** Eric Brynjolfsson, Andrew McAfee. The second machine age. Norton & Company. 2014
90. **Bryson & Ho (1969)** Bryson, A. and Ho, Y. (1969). Applied optimal control. Blaisdell, Waltham, MA.
91. **Bubeck et al. (2023)** Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang. Sparks of Articial General Intelligence: Early experiments with GPT-4. arXiv: 2303.12712v3 [cs.CL] 27 Mar 2023
92. **Budson et al. (2022)** Andrew E. Budson, Kenneth A. Richman, and Elizabeth A. Kensinger, Consciousness as a Memory System, Cogn Behav Neurol 2022
93. **Buehler (2024)** Markus J. Buehler. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. arXiv:2403.11996v2 [cs.LG] 26 Mar 2024
94. **Butlin et al. (2023)** Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708v3 [cs.AI] 22 Aug 2023
95. **CAICT (2021)** White Paper on Trustworthy Artificial Intelligence 可信人工智能白皮书, China, 2021
96. **CAICT (2022)** Artificial Intelligence White Paper (2022), 人工智能白皮书（2022年), The China Academy of Information and Communications Technology, CAICT website, April 12, 2022.
97. **Caldarelli (2007)** G. Caldarelli, Scale-free networks: complex webs in nature and technology (Oxford University Press, 2007).
98. **Cantlon & Piantadosi (2024)** Jessica F. Cantlon & Steven T. Piantadosi. Uniquely human intelligence arose from expanded information capacity. Nature reviews psychology. Perspective. 2024
99. **Carvalho & Damasio (2021)** Carvalho, G. B. & Damasio, A. Interoception and the origin of feelings: a new synthesis. Bioessays 43, e2000261 (2021)
100. **CB (2023)** CB Insights. Generative AI Bible: The ultimate guide to genAI disruption. Researh Report. https://www.cbinsights.com/research/report/generative-ai-bible/, November 7, 2023
101. **Chae et al. (2024)** Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, Jinyoung Yeo. Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation. arXiv:2410.13232v1 [cs.CL] 17 Oct 2024
102. **Chang (2025a)** Edward Y. Chang. Multi-LLM Agent Collaborative Intelligence: The Path to AGI. Imprint: SocraSynth.com (2025)
103. **Chang (2025b)** Edward Y. Chang. MACI: Multi-Agent Collaborative Intelligence for Adaptive Reasoning and Temporal Planning. MACI Version 1: January 26, 2025, Stanford University

104. **Chang et al. (2020)** Chang, A. Y. C., Biehl, M., Yu, Y. & Kanai, R. Information closure theory of consciousness. Front. Psychol. 11, 1504 (2020)

105. **Chen & Li (2024)** Wei Chen and Zhiyuan Li. Octopus v4: Graph of language models. arXiv:2404.19296v1 [cs.CL] 30 Apr 2024

106. **Chen et al. (2024)** Justin Chih-Yao Chen, Zifeng Wang, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee, Tomas Pfister. Reverse Thinking Makes LLMs Stronger Reasoners. arXiv:2411.19865v1 [cs.CL] 29 Nov 2024

107. **Cheng et al. (2024)** Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, Nan Du. Self-playing Adversarial Language Game Enhances LLM Reasoning. arXiv:2404.10642v1 [cs.CL] 16 Apr 2024

108. **Chollet (2019)** François Chollet. The measure of intelligence. arXiv preprint arXiv:1911.01547, 2019.

109. **Chomsky (1957)** Chomsky, N. Syntactic Structures. The Hague: Mouton. 1957

110. **Chomsky et al. (1976)** Noam Chomsky et al. Reflections on language. Temple Smith London, 1976.

111. **Chomsky et al. (2006)** Noam Chomsky et al. Language and Mind. Cambridge University Press, 2006.

112. **Chowdhery et al. (2022)** Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., et al. PaLM: Scaling language modeling with Pathways. arXiv preprint arXiv:2204.02311, 2022.

113. **Christakopoulou, Mourad & Matari´c (2024)** Konstantina Christakopoulou, Shibl Mourad, Maja Matari´c. Agents Thinking Fast and Slow: A Talker-Reasoner Architecture. arXiv:2410.08328v1 [cs.AI] 10 Oct 2024

114. **Clark (2013)** Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36, 181–204 (2013)

115. **Clay, Leadholm & Hawkins (2024)** Viviane Clay, Niels Leadholm, and Jeff Hawkins. The Thousand Brains Project. Numenta, 2024

116. **Cleeremans (2021)** Cleeremans, A. The radical plasticity thesis: how the brain learns to be conscious. Front. Psychol. 2, 86 (2011)

117. **Cleeremans et al. (2020)** Cleeremans, A. et al. Learning to be conscious. Trends Cogn. Sci. 24, 112–123 (2020)

118. **CNAS (2023)** Jacob Stokes and Alexander Sullivan with Noah Greene. U.S.-China Competition and Military AI. How Washington Can Manage Strategic Risks amid Rivalry with Beijing. CNAS. JULY 2023

119. **Cohen & Havlin (2010)** R. Cohen and S. Havlin, Complex networks: structure, robustness and function (Cambridge University Press, 2010).

120. **Conant & Ashby (1970)** Roger C. Conant and W. Ross Ashby. “Every good regulator of a system must be a model of that system”. In: International Journal of Systems Science 1.2 (1970), pp. 89–97.

121. **Constant, Friston & Clark (2023)** Constant A, Friston KJ, Clark A. Cultivating creativity: predictive brains and the enlightened room problem. Phil. Trans. R. Soc. B 379: 20220415. (2023)

122. **Copeland (2000)** Jack Copeland. What is Artificial Intelligence? © Copyright B.J. Copeland, May 2000 http://www.alanturing.net/turing_archive/pages/Reference Articles/What is AI.html

123. **CRFM (2021)** On the Opportunities and Risks of Foundation Models. Authored by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI) 2021

124. **Cross et al. (2024)** Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, Nick Haber. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. arXiv:2407.07086v1 [cs.AI] 9 Jul 2024

125. **Da Costa et al. (2020)** Da Costa, L. et al. Active inference on discrete state-spaces: A synthesis. Journal of Mathematical Psychology 99, 102447, (2020).

126. **Da Costa et al. (2024a)** Lancelot Da Costa, Tomáš Gavenčiak, David Hyland, Mandana Samiei, Cristian Dragos-Manta, Candice Pattisapu, Adeel Razi, Karl Friston. Possible principles for aligned structure learning agents. arXiv:2410.00258v1 [cs.AI] 30 Sep 2024

127. **Da Costa et al. (2024b)** Lancelot Da Costa, Lars Sandved-Smith, Karl Friston, Maxwell J. D. Ramstead, Anil K. Seth. A Mathematical Perspective on Neurophenomenology. arXiv:2409.20318v1 [q-bio.NC] 30 Sep 2024

128. **Dafoe et al. (2020)** Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. Open problems in cooperative AI. arXiv preprint arXiv:2012.08630.

129. **Dalrymple et al. (2024)** Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., ... & Tenenbaum, J. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. arXiv preprint arXiv:2405.06624.

130. **Damasio (2000)** Damasio, A. The Feeling of What Happens: Body and Emotion in the Making of Consciousness (Harvest Books, 2000)

131. **Damasio (2010)** Damasio, A. Self Comes To Mind: Constructing the Conscious Brain (William Heinemann, 2010)

132. **Danilenka et al. (2024)** Anastasiya Danilenka, Alireza Furutanpey, Victor Casamayor Pujol, Boris Sedlak, Anna Lackinger, Maria Ganzha, Marcin Paprzycki, Schahram Dustdar. Adaptive Active Inference Agents for Heterogeneous and Lifelong Federated Learning. arXiv:2410.09099v1 [cs.LG] 9 Oct 2024

133. **De Chardin (1999)** Pierre Teilhard de Chardin, The human phenomenon; a new edition and translation of Le phénomène humain by Sarah Appleton-Weber; with a foreword by Brian Swimme Brighton [UK] ; Portland, Or. : Sussex Academic Press, (1999)

134. **Deane (2021)** Deane, G. Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution. Neuroscience of Consciousness, 2021(2), niab024.

135. **DeepMind Adaptive Agents Team (2023)** AdA Team – Adaptive Agents Model. Deep Mind.

136. **DeepSeek (2025)** DeepSeek-AI (Team). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948v1 [cs.CL] 22 Jan 2025

137. **Dehaene & Changeux (2011)** Dehaene, S., & Changeux, J. P. Experimental and theoretical approaches to conscious processing. Neuron; 70(2), pp. 200-227.

138. **Dehaene & Naccache (2001)** Dehaene, S., & Naccache, L. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79, 2001, pp.1–37.

139. **Dehaene (2014)** S. Dehaene, Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts (Viking Press, New York, NY, 2014).

140. **Dehaene et al. (1998)** Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the national Academy of Sciences, 95(24):14529–14534, 1998.

141. **Dehaene et al. (2003**) Dehaene, S., Sergent, C., & Changeux, J. P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Sciences, 100, 2003, 8520–8525.

142. **Dehaene et al. (2006)** Stanislas Dehaene, Jean-Pierre Changeux, Lionel Naccache, Jérôme Sackur, and Claire Sergent. Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in cognitive sciences, 10(5) 2006:204–211.

143. **Dehaene et al. (2021)** Dehaene, S., Lau, H., and Kouider, S. What is consciousness, and could machines have it? Robotics, AI, and Humanity, pages 43-56.

144. **Dehaene et al. (2022)** Dehaene S., Al Roumi F., Lakretz Y., Planton S. & Sablé-Meyer M. Symbols and mental programs: a hypothesis about human singularity. Trends in Cognitive Sciences. 2022

145. **Delphi (2021)** Delphi: Towards machine ethics and norms. Paul G. Allen School of Computer Science & Engineering, University of Washington Allen Institute for Artificial Intelligence, 2021

146. **Dennett (1991)** Dennett, D. C. Consciousness Explained (Little, Brown, 1991)

147. **DGA-ASG (2024)** AI Decrypted: A Guide for Navigating AI. Developments in 2024. DENTONS GLOBAL ADVISORS - ALBRIGHT STONEBRIDGE GROUP. 2024

148. **Didolkar et al. (2024)** Didolkar, A., Goyal, A., Ke, N. R., Guo, S., Valko, M., Lillicrap, T., ... & Arora, S. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. arXiv:2405.12205.

149. **Dolan & Dayan (2013)** Dolan, R. J., & Dayan, P. Goals and habits in the brain. Neuron, 80(2), 2013, pp.312–325.

150. **Dorogovtsev & Mendes (2003)** S. N. Dorogovtsev and J. F. Mendes, Evolution of networks: From biological nets to the Internet and WWW (Oxford University Press, 2003).

151. **Driess et al. (2023)** Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., & Florence, P. PaLM-E: An embodied multimodal language model., arXiv:2303.03378, 2023

152. **Durante et al. (2024)** Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, Demetri Terzopoulos, Ade Famoti, Noboru Kuno, Ashley Llorens, Hoi Vo, Katsu Ikeuchi, Li Fei-Fei, Jianfeng Gao, Naoki Wake, Qiuyuan Huang. An Interactive Agent Foundation Model. arXiv:2402.05929v1 [cs.AI] 8 Feb 2024

153. **Edelman (1987)** Edelman, G. M. Neural Darwinism: The Theory of Neuronal Group Selection (Basic Books 1987).

154. **Edelman (1989)** Edelman, G. M. The Remembered Present (Basic Books, 1989)

155. **Elhage et al. (2021)** Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, et al. “A Mathematical Framework for Transformer Circuits”. In: Transformer Circuits Thread (2021).

156. **Feng et al. (2024)** Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, Jiaxuan You. How Far Are We From AGI? arXiv:2405.10313v1 [cs.AI] 16 May 2024

157. **Ferreira, Silva & Martins (2024)** Silvan Ferreira, Ivanovitch Silva, and Allan Martins. Organizing a Society of Language Models: Structures and Mechanisms for Enhanced Collective Intelligence. arXiv:2405.03825 [cs.AI]

158. **Fields, Glazebrook & Levin (2024)** Fields, C.; Glazebrook, J.F.; Levin, M. Principled Limitations on Self-Representation for Generic Physical Systems. Entropy 2024, 26, 194.

159. **Flake (2006)** Gary William Flake. How i learned to stop worrying and love the imminent internet singularity. In Proceedings of the 15th ACM international conference on Information and knowledge management, page 2, 2006. Arlington, Virginia, USA.

160. **Fleming (2020)** Fleming, S. M. Awareness as inference in a higher-order state space. Neuroscience of consciousness, 2020, 020.

161. **Francis & Wonham (1976)** Bruce A. Francis and Walter M. Wonham. “The internal model principle of control theory”. In: Automatica 12.5 (1976), pp. 457–465.

162. **Frank et al. (2022)** Adam Frank, David Grinspoon, and Sara Walker. “Intelligence as a planetary scale process”. In: International Journal of Astrobiology 21.2 (2022), pp. 47–61.

163. **Friston (2010)** Karl Friston. The free-energy principle: a unified brain theory? Nature reviews neuroscience, 11(2) 2010:127–138.

164. **Friston (2013)** Karl Friston. “Life as we know it”. In: Journal of the Royal Society Interface 10.86 (2013), p. 20130475.
165. **Friston (2018)** Friston, K. J. Am I self-conscious? (Or does self-organization entail self-consciousness?). Front. Psychol. 9, 579 (2018)
166. **Friston (2019)** Karl Friston. “A free energy principle for a particular physics”. In: arXiv (2019).
167. **Friston et al. (2015)** Karl Friston, Michael Levin, Biswa Sengupta, and Giovanni Pezzulo. “Knowing one’s place: a free-energy approach to pattern regulation”. In: Journal of The Royal Society Interface 12.105 (2015).
168. **Friston et al. (2017)** Karl Friston, Thomas Parr, and Bert de Vries. “The graphical brain: Belief propagation and active inference”. In: Network Neuroscience 1.4 (2017), pp. 381–414.
169. **Friston et al. (2020)** Karl Friston, Thomas Parr, Yan Yufik, Noor Sajid, Catherine J. Price, and Emma Holmes. “Generative models, linguistic communication and active inference”. In: Neuroscience and Biobehavioral Reviews 118 (2020), pp. 42–64.
170. **Friston et al. (2021)** Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. “Sophisticated Inference”. In: Neural Computation 33.3 (2021), pp. 713–763.
171. **Friston et al. (2022)** Karl J. Friston, Maxwell J.D. Ramstead, Alex B. Kiefer, Alexander Tschantz, Christopher L. Buckley, Mahault Albarracin, Riddhi J. Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton A.R. Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G. Fox, Steven Swanson, Dan Mapes, and Gabriel René. Designing Ecosystems of Intelligence from First Principles. arXiv:2212.01354v1 [cs.AI] 2 Dec 2022
172. **Friston et al. (2023)** Karl J. Friston, Tommaso Salvatori, Takuya Isomura, Alexander Tschantz, Alex Kiefer, Tim Verbelen, Magnus Koudahl, Aswin Paul, Thomas Parr, Adeel Razi, Brett Kagan, Christopher L. Buckley, and Maxwell J. D. Ramstead. Active Inference and Intentional Behaviour. arXiv:2312.07547v2 [q-bio.NC] 16 Dec 2023
173. **Fudan (2023)** Fudan NLP Group. The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv:2309.07864v3 [cs.AI] 19 Sep 2023
174. **Ginsburg & Jablonka (2019)** Ginsburg, S. & Jablonka, E. The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness (MIT Press, 2019)
175. **Glück & Bluck (2013)** Glück, J., & Bluck, S. The MORE Life Experience Model: A theory of the development of personal wisdom. In M. Ferrari & N. M. Weststrate (Eds.), The scientific study of personal wisdom (pp. 75–98). Berlin, Germany: Springer. (2013)
176. **Glück & Weststrate (2022)** Glück, J., & Weststrate, N. M. The wisdom researchers and the elephant: An integrative model of wise behavior. Personality and Social Psychology Review, 26, 342–374. (2022)
177. **Glück et al. (2005)** Glück, J., Bluck, S., Baron, J., & McAdams, D. The wisdom of experience: Autobiographical narratives across adult- hood. International Journal of Behavioral Development, 29, 197–208. (2005)
178. **Godfrey-Smith (2016)** Godfrey-Smith, P. Mind, matter, and metabolism. The Journal of Philosophy, 113, 2016, pp.481– 506.
179. **Godfrey-Smith (2019)** Godfrey-Smith, P. Evolving across the explanatory gap. Philosophy, Theory, and Practice in Biology, 11(1), 2019.
180. **Goertzel (2006)** Ben Goertzel. The Hidden Pattern. Brown Walker, 2006.
181. **Goertzel (2014a)** Ben Goertzel. The AGI Revolution. Amazon, 2014.
182. **Goertzel (2014b)** Ben Goertzel. Golem: towards an agi meta-architecture enabling both goal preservation and radical self-improvement. J. Exp. Theor. Artif. Intell., 26(3):391–403, 2014.

183. **Goertzel (2017a)** Ben Goertzel. Toward a formal model of cognitive synergy. CoRR, abs/1703.04361, 2017.

184. **Goertzel (2017b)** Ben Goertzel. Euryphysics: a (somewhat) new conceptual model of mind, reality and psi. Journal of Nonlocality, 5(1), 2017.

185. **Goertzel (2019)** Ben Goertzel. Distinction graphs and graphtropy: A formalized phenomenological layer underlying classical and quantum entropy, observational semantics and cognitive computation. CoRR, abs/1902.00741, 2019.

186. **Goertzel (2020)** Ben Goertzel. Paraconsistent foundations for probabilistic reasoning, programming and concept formation. arXiv preprint arXiv:2012.14474, 2020.

187. **Goertzel (2021a)** Ben Goertzel. The General Theory of General Intelligence: A Pragmatic Patternist Perspective. arXiv:2103.15100v3 [cs.AI] 4 Apr 2021

188. **Goertzel (2021b)** Ben Goertzel. Patterns of cognition: Cognitive algorithms as galois connections fulfilled by chronomorphisms on probabilistically typed metagraphs. arXiv preprint arXiv:2102.10581, 2021.

189. **Goertzel et al. (2023)** Ben Goertzel, Vitaly Bogdanov, Michael Duncan, Deborah Duong, Zarathustra Goertzel, Jan Horlings, Matthew Ikle', Lucius Greg Meredith, Alexey Potapov, Andre' Luiz de Senna, Hedra Seid Andres Suarez, Adam Vandervorst, Robert Werko. OpenCog Hyperon: A Framework for AGI at the Human Level and Beyond. arXiv:2310.18318v1 [cs.AI] 19 Sep 2023

190. **Goertzel, Ikle' & Wigmore (2012)** Ben Goertzel, Matt Ikle', and Jared Wigmore. The architecture of human-like general intelligence. In Foundations of Artificial General Intelligence, 2012.

191. **Goertzel, Pennachin & Geisweiller (2013a)** Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy. Springer: Atlantis Thinking Machines, 2013.

192. **Goertzel, Pennachin & Geisweiller (2013b)** Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI. Springer: Atlantis Thinking Machines, 2013.

193. **Good (1965)** I.J. Good. Speculations concerning the first ultraintelligent machine. Advances in Computers, 6, 1965.

194. **Google (2023)** PaLM 2 Technical Report. Google, 2023

195. **Google DeepMind (2023a)** Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, Tim Rockt¨aschel. PROMPTBREEDER: self-referential self-improvement via prompt evolution. arXiv:2309.16797v1 [cs.CL] 28 Sep 2023

196. **Google DeepMind (2023b)** Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet and Shane Legg. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462v1 [cs.AI] 4 Nov 2023

197. **Google DeepMind (2023c)** Google DeepMind. Welcome to the Gemini era. https://deepmind.google/technologies/gemini/#introduction

198. **Google DeepMind (2023d)** Google DeepMind. Gemini Team. Gemini: A Family of Highly Capable Multimodal Models. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

199. **Google DeepMind (2024a)** Google DeepMind. DiPaCo: Distributed Path Composition. arXiv:2403.10616v1 [cs.LG] 15 Mar 2024

200. **Google DeepMind (2024b)** Google DeepMind. Evaluating Frontier Models for Dangerous Capabilities. arXiv:2403.13793v1 [cs.LG] 20 Mar 2024

201. **Google DeepMind (2024c)** Google DeepMind. Long-form factuality in large language models. arXiv:2403.18802v1 [cs.CL] 27 Mar 2024

202. **Graziano & Webb (2015)** Michael S. A. Graziano and Taylor W Webb. The attention schema theory: a mechanistic account of subjective awareness. Frontiers in psychology, 6:500, 2015.
203. **Graziano (2017)** Graziano, M. S. A. The attention schema theory: a foundation for engineering artificial consciousness. Front. Robot. AI 4, 60 (2017)
204. **Graziano (2019)** Graziano, M. S. Rethinking Consciousness: A Scientific Theory of Subjective Experience. WW Norton & Company. 2019
205. **Graziano et al. (2020)** Michael S. A. Graziano, Arvid Guterstam, Branden J Bio, and Andrew I Wilterson. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. Cognitive Neuropsychology, 37(3-4):155–172, 2020.
206. **Greyling (2024)** Cobus Greyling. Five Levels Of AI Agents. https://cobusgreyling.medium.com/five-levels-of-ai-agents-5ac39a7b07ed May 16, 2024
207. **Grossmann (2017)** Grossmann, I. Wisdom in context. Perspectives on Psychological Science, 12, 233–257. (2017)
208. **Grossmann et al. (2020)** Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., ... & Vervaeke, J. The science of wisdom in a polarized world: Knowns and unknowns. Psychological Inquiry, 31, 103–133. (2020)
209. **Gu et al. (2024)** Yu Gu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, Yu Su. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning For Web Agents. arXiv:2411.06559v1 [cs.AI] 10 Nov 2024
210. **Gurnee & Tegmark (2023)** Wes Gurnee & Max Tegmark. Language Models Represent Space and Time. arXiv:2310.02207v1 [cs.LG] 3 Oct 2023
211. **Hadsell et al. (2020)** Hadsell, R., Rao, D., Rusu, A. A., Pascanu, R. Embracing change: Continual learning in deep neural networks. Trends in Cognitive Sciences, 24 (12) 2020:1028-1040.
212. **Haken & Haken-Krell (1994)** Haken, Hermann, Haken-Krell, Maria. Erfolgsgeheimnisse der Wahrnehmung. Synergetik als Schlüssel zum Gehirn. Die schillernde Welt der Gehirn- und Computerforschung. Frankfurt, Berlin. Ullstein 1994
213. **Haken (1978)** Haken, Hermann. Synergetics: an introduction: nonequilibrium phase transitions and self-organization in physics, chemistry, and biology. Berlin New York: Springer-Verlag, 1978
214. **Hakenes & Irmen (2004)** Hendrik Hakenes and Andreas Irmen. Airy growth was the take-off inevitable? 2004.
215. **Hakenes & Irmen (2007)** Hendrik Hakenes and Andreas Irmen. On the longrun evolution of technological knowledge. Economic Theory, 30:171-180, 2007.
216. **Hameroff & Penrose (2014)** Hameroff, S. & Penrose, R. Consciousness in the universe: a review of the 'Orch OR' theory. Phys. Life Rev. 11, 39–78 (2014)
217. **Hammacher (2006)** Kay Hammacher. Accelerating changes in our epoch and the role of time-horizons. In Vladimir Burdyuzha, editor, The Future of Life and the Future of our Civilization, volume III. Springer, 2006.
218. **Hanson (1998a)** Robin Hanson. Economic growth given machine intelligence. http://hanson.gmu.edu/aigrow.pdf, 1998.
219. **Hanson (1998b)** Robin Hanson. Is a singularity just around the corner? what it takes to get explosive economic growth. Journal of Evolution and Technology, 2, 1998. http://hanson.gmu.edu/fastgrow.html.
220. **Hanson (1998c)** Robin Hanson. Long-term growth as a sequence of exponential modes. http://hanson.gmu.edu/longgrow.pdf, 1998.

221. **Hanson (2008a)** Robin Hanson. Economics of brain emulations. In Peter Healey and Steve Rayner, editors, Unnatural Selection - The Challenges of Engineering Tomorrow's People,, pages 150-158. EarthScan, London, 2008.

222. **Hanson (2008b)** Robin Hanson. Economics of the singularity. IEEE Spectrum, pages 37-42, June 2008.

223. **Hawkins, Jeff & Ahmad (2016)** “Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex”. In: Frontiers in Neural Circuits 10. issn: 16625110.

224. **Hawkins, Jeff et al. (2019)** Hawkins, Jeff et al. “A framework for intelligence and cortical function based on grid cells in the neocortex”. In: Frontiers in Neural Circuits. issn: 16625110.

225. **Hawkins, Jeff, Ahmad & Cui (2017)** Hawkins, Jeff, Ahmad and Cui “A Theory of How Columns in the Neocortex Enable Learning the Structure of the World”. In: Frontiers in Neural Circuits 11.October, pp. 1–18. issn: 1662-5110.

226. **Hendrycks (2023)** Dan Hendrycks. Natural Selection Favors AIs over Humans. arXiv:2303.16200v1 [cs.CY] 28 Mar 2023

227. **Henriques et al. (2019)** Henriques, G., Michalski, J., Quackenbush, S., Schmidt, W. The Tree of Knowledge System: A New Map for Big History. Journal of Big History, III (4); 1 - 17.

228. **Hesp et al. (2020)** Casper Hesp, Alexander Tschantz, Beren Millidge, Maxwell Ramstead, Karl Friston, and Ryan Smith. “Sophisticated affective inference: simulating anticipatory affective dynamics of imagining future events”. In: International Workshop on Active Inference. Springer. 2020, pp. 179–186.

229. **Heylighen (1997)** Francis Heylighen. The socio-technological singularity. http://pespmc1.vub.ac.be/SINGULAR.html, 1997

230. **Heylighen (2007)** Francis Heylighen. Accelerating socio-technological evolution: from ephemeralization and stigmergy to the global brain. In George Modelski, Tessaleno Devezas and William Thompson, editors, Globalization as an Evolutionary Process: Modeling Global Change. Routledge, London, 2007

231. **Ho et al. (2023)** Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, Duncan Snidal. International Institutions for Advanced AI. arXiv:2307.04699v2 [cs.CY] 11 Jul 2023

232. **Hochberg (2024)** Michael E. Hochberg. A Theory of Intelligences. Preprints.org. 202404.0722.v2. 27 May 2024

233. **Hoffman & GPT-4 (2023)** Reid Hoffman with GPT-4. Impromptu. Amplifying Our Humanity Through AI. Dallepedia LLC, 2023

234. **Hoffmann et al. (2022)** Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., et al. Training compute-optimal large language models. NeurIPS, 2022.

235. **Hohwy & Seth (2020)** Hohwy, J. & Seth, A. K. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Philos. Mind Sci. 1, 3 (2020)

236. **Hohwy (2013)** Hohwy, J. The Predictive Mind (Oxford Univ. Press, 2013)

237. **Hohwy (2022)** Hohwy, J. Conscious self-evidencing. Review of Philosophy and Psychology, 13(4), 2022, pp.809–828.

238. **Holt (2024)** Denise Holt. VERSES AI's Active Inference Outperforms Deep Learning in Historic AI Industry Benchmark Test. https://deniseholt.substack.com/p/verses-ai-active-inference-beats-deep-learning 9 Mar 2024

239. **Hoyle (2024)** Victoria Violet Hoyle. The Phenomenology of Machine: A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures. arXiv:2410.00033v1 [cs.AI] 18 Sep 2024

240. **Hu & Clune (2023)** Shengran Hu &TJeff Clune. Thought Cloning: Learning to Think while Acting by Imitating Human Thinking. arXiv:2306.00323v1 [cs.AI] 1 Jun 2023

241. **Huang (2024)** Yu Huang. Levels of AI Agents: from Rules to Large Language Models. arXiv:2405.06643 [cs.CL] 6 Mar 2024

242. **Huang et al. (2021)** Huang, H.-Y. et al. Power of data in quantum machine learning. Nat. Commun. 12, 1 (2021)

243. **Huh et al. (2024)** Minyoung Huh, Brian Cheung, Tongzhou Wang, Phillip Isola. The Platonic Representation Hypothesis. arXiv:2405.07987v1 [cs.LG] 13 May 2024

244. **Isomura et al. (2023)** Takuya Isomura, Kiyoshi Kotani, Yasuhiko Jimbo, Karl J. Friston. Experimental validation of the free-energy principle with in vitro neural networks. Nature Communications 14, 4547 (2023).

245. **Jackendoff (1987)** Jackendoff, R. Consciousness and the Computational Mind (MIT Press, 1987).

246. **Jaegle et al. (2021a)** Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. Perceiver: General perception with iterative attention. International Conference on Machine Learning. PMLR, 2021, pp.4651–4664.

247. **Jaegle et al. (2021b)** Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E. Perceiver IO: A general architecture for structured inputs & outputs. arXiv:2107.14795. 2021

248. **Jakobson (1965)** Roman Jakobson. Quest for Essence of Language / «Diogenes. An International Review of Philosophy and Humanistic Studies», Montreal, 1965, № 51, с. 21—37.

249. **Johansen & Sornette (2001)** Anders Johansen and Didier Sornette. Finite-time singularity in the dynamics of the world population, economic and financial indices. Physica A, 294:465-502, 2001.

250. **Johnson et al. (2024)** Samuel G. B. Johnson, Amir-Hossein Karimi, Yoshua Bengio, Nick Chater, Tobias Gerstenberg, Kate Larson, Sydney Levine, Melanie Mitchell, Iyad Rahwan, Bernhard Schölkopf, Igor Grossmann. Imagining and building wise machines: The centrality of AI metacognition. arXiv:2411.02478v1 [cs.AI] 4 Nov 2024

251. **Jones & Bergen (2023)** Cameron Jones and Benjamin Bergen. Does GPT-4 Pass the Turing Test? arXiv:2310.20216v1 [cs.AI] 31 Oct 2023

252. **Juliani et al. (2022)** Arthur Juliani, Kai Arulkumaran, Shuntaro Sasai, Ryota Kanai. On the link between conscious function and general intelligence in humans and machines. arXiv:2204.05133v2 [cs.AI] 19 Jul 2022.

253. **Kadavath et al. (2022)** Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, Jared Kaplan. Language Models (Mostly) Know What They Know. arXiv:2207.05221v3 [cs.CL] 16 Jul 2022

254. **Kahneman (2011)** Kahneman, D. Thinking, fast and slow. Macmillan, 2011

255. **Kak & West (2023)** Amba Kak and Sarah Myers West, “AI Now 2023 Landscape: Confronting Tech Power”, AI Now Institute, April 11, 2023, https://ainowinstitute.org/2023-landscape.

256. **Kanai et al. (2019)** Ryota Kanai, Acer Chang, Yen Yu, Ildefons Magrans de Abril, Martin Biehl, and Nicholas Guttenberg. Information generation as a functional basis of consciousness. Neuroscience of Consciousness, 2019(1): niz016, 2019.
257. **Kaplan et al. (2020)** Kaplan, J. et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020).
258. **Kephart & Chess (2003)** Jeff Kephart and David Chess. The vision of autonomic computing. Computer 36, 1 (Jan 2003), 41–50.
259. **Khattab et al. (2023)** Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, Christopher Potts. DSPY: Compiling Declarative Language Model Calls into Self-Improving Pipelines. arXiv:2310.03714v1 [cs.CL] 5 Oct 2023
260. **Khetarpal et al. (2020)** Khetarpal, K., Riemer, M., Rish, I., Precup, D. Towards continual reinforcement learning: A review and perspectives, arXiv:2012.13490, 2020
261. **Kim et al. (2024)** Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, Minjoon Seo. PROMETHEUS 2: An Open Source Language Model Specialized in Evaluating Other Language Models. arXiv:2405.01535v1 [cs.CL] 2 May 2024
262. **Koppl et al (2021)** Koppl, Roger and Devereaux, Abigail and Valverde, Sergi and Solé, Ricard and Kauffman, Stuart and Herriot, James, Explaining Technology (May 30, 2021). Available at SSRN: https://ssrn.com/abstract=3856338
263. **Kornieiev (2025)** Sergii Kornieiev. Cognitive architecture AGICA: “Space of Reasoning of individual common sense”. Preprint · February 2025 DOI: 10.31219/osf.io/ct5jz_v1
264. **Kosinski (2023)** Michal Kosinski. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv:2302.02083v2
265. **Kuhn (2024)** R. L. Kuhn, “A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications,” Progress in Biophysics and Molecular Biology, 190, 28–169 [Aug.] (2024)
266. **Kurzweil (2001)** Raymond Kurzweil. The law of accelerating returns. http://www.kurzweilai.net/articles/art0134.html, March 7 2001.
267. **Kurzweil (2005)** Raymond Kurzweil. The Singularity Is Near: When Humans Transcend Biology. Viking Penguin, 2005.
268. **Kurzweil (2012)** Kurzweil, Ray, How to Create a Mind: The Secret of Human Thought Revealed, New York: Viking Books, 2012
269. **KVM** Ray Kurzweil, Vernor Vinge, and Hans Moravec. Singularity math trialogue. http://www.kurzweilai.net/meme/frame.html?main=/articles/art0151.html.
270. **Lahav & Neemeh (2022)** Lahav N and Neemeh ZA A Relativistic Theory of Consciousness. Front. Psychol. 12:704270. (2022)
271. **Lake et al. (2017)** Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. “Building machines that learn and think like people”. en. In: Behav. Brain Sci. 40 (2017), e253.
272. **Lamme (2006)** Lamme, V. A. F. Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 2006, pp.494–501.
273. **Lamme (2010)** Lamme, V. A. F. How neuroscience will change our view on consciousness. Cognitive Neuroscience, 1(3), 2010, pp.204–220.
274. **Lamme (2020)** Lamme, V. A. F. Visual functions generate conscious seeing. Frontiers in Psychology, 11, 83, 2020
275. **Lau (2019)** Lau, H. Consciousness, metacognition, & perceptual reality monitoring. PsyArXiv. 2019

276. **Lau (2022)** Lau, H. In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience. Oxford University Press. 2022

277. **LeCun (2022)** Yann LeCun. A Path Towards Autonomous Machine Intelligence. Version 0.9.2, 2022-06-27, Courant Institute of Mathematical Sciences, New York University. Meta - Fundamental AI Research, 2022

278. **LeCun et al. (2006)** LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. In Bakir, G., Hofman, T., Schoelkopf, B., Smola, A., and Taskar, B., editors, Predicting Structured Data. MIT Press. (2006)

279. **Lee (2022)** Lee, A. Y. Degrees of consciousness. Noˆus. 2022

280. **Lee et al. (2021)** Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499, 2021.

281. **Lee et al. (2024)** Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. MoAI: Mixture of All Intelligence for Large Language and Vision Models. arXiv:2403.07508v1 [cs.CV] 12 Mar 2024

282. **Legg (2022)** Shane Legg. Twitter (now "X"), May 2022. URL https://twitter.com/ShaneLegg/status/1529483168134451201. Accessed on October 12, 2023.

283. **Leontief (1986)** Wassily W. Leontief. Input-output economics. Oxford University Press, 2nd edition, 1986.

284. **Levin (2024)** Michael Levin. Self-Improvising Memory: A Perspective on Memories as Agential, Dynamically Reinterpreting Cognitive Glue. Entropy 2024, 26, 481.

285. **Lewis, Marcus et al. (2019)** “Locations in the neocortex: A theory of sensorimotor object recognition using cortical grid cells”. In: Frontiers in Neural Circuits. issn: 16625110.

286. **Li et al. (2022)** Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d’Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models. In Empirical Methods in Natural Language Processing, pages 11838–11855, 2022.

287. **Li et al. (2023)** Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, Xing Xie. Large Language Models Understand and Can Be Enhanced by Emotional Stimuli. arXiv:2307.11760v6 [cs.CL] 6 Nov 2023

288. **Li et al. (2024a)** Jialong Li, Mingyue Zhang, Nianyu Li, Danny Weyns, Zhi Jin, and Kenji Tei. Generative AI for Self-Adaptive Systems: State of the Art and Research Roadmap. Article in ACM Transactions on Autonomous and Adaptive Systems · September 2024

289. **Li et al. (2024b)** Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu and Yi Yang. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth (2024) 1:9

290. **Liang et al. (2024)** Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, James Y. Zou. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. arXiv:2403.07183v1 [cs.CL] 11 Mar 2024

291. **Lightman et al. (2023)** Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, Karl Cobbe. Let’s Verify Step by Step. arXiv:2305.20050v1 [cs.LG] 31 May 2023

292. **Lin et al. (2023)** Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2Motion: From natural language instructions to feasible plans. arXiv preprint arXiv:2303.12153, 2023.

293. **Liu et al. (2021)** Y. Liu, N. Dehmamy, and A.-L. Baraba'si, Isotopy and energy of physical networks, Nature Physics 17, 216 (2021).
294. **Liu et al. (2023)** Liu, D., Bolotta, S., Zhu, H., Bengio, Y., & Dumas, G. Attention schema in neural agents. arXiv:2305.17375. 2023
295. **Liu et al. (2024)** Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljaˇci´c, Thomas Y. Hou, Max Tegmark. KAN: Kolmogorov–Arnold Networks. arXiv:2404.19756v2 [cs.LG] 2 May 2024
296. **Lobo & Del Ser (2024)** Jesús López Lobo and Javier Del Ser. Can transformative AI shape a new age for our civilization?: Navigating between speculation and reality. arXiv:2412.08273v1 [cs.AI] 11 Dec 2024
297. **Luppi et al. (2024)** Andrea I. Luppi, Fernando E., Rosas Pedro, A.M. Mediano, David K. Menon, Emmanuel A. Stamatakis. Information decomposition and the informational architecture of the brain. Trends in Cognitive Sciences, 2023. Published: January 09, 2024.
298. **Ma S. et al. (2024)** Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, Furu Wei. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv:2402.17764v1 [cs.CL] 27 Feb 2024
299. **Ma X. et al. (2024)** Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, Chunting Zhou. MEGALODON: Efficient LLM Pretraining and Inference with Unlimited Context Length. arXiv:2404.08801v2 [cs.LG] 16 Apr 2024
300. **MAD (2023)** The 2023 MAD (ML/AI/Data) Landscape. https://mad.firstmark.com/
301. **Man & Damasio (2019)** Man, K., & Damasio, A. Homeostasis and soft robotics in the design of feeling machines. Nature Machine Intelligence, 1(10), 2019, pp.446–452.
302. **Manzano et al. (2024)** Gonzalo Manzano , Gülce Kardeş, Édgar Roldán, and David H. Wolpert. Thermodynamics of Computations with Absolute Irreversibility, Unidirectional Transitions, and Stochastic Computation Times. PHYSICAL REVIEW X 14, 021026 (2024)
303. **Marcus (2001)** Gary Marcus. The algebraic mind, 2001.
304. **Marcus (2023a)** Gary Marcus. Dear Elon Musk, here are five things you might want to consider about AGI. "Marcus on AI" Substack, May 2022. URL https://garymarcus.substack.com/p/dear-elon-musk-here-are-five-things?s=r.
305. **Marcus (2023b)** Gary Marcus. Twitter (now "X"), May 2022. URL https://twitter.com/GaryMarcus/status/1529457162811936768. Accessed on October 12, 2023.
306. **Marr (1982)** D Marr. Vision: A computational approach. Freeman & Co., San Francisco, 1982.
307. **Mashour et al. (2020)** Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. Neuron, 105, 2020, pp.776–798.
308. **Maslej et al. (2023)** Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.
309. **Masterman et al. (2024)** Tula Masterman, Sandi Besen, Mason Sawtell, Alex Chao. The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: a survey. arXiv:2404.11584v1 [cs.AI] 17 Apr 2024
310. **Mazzaglia et al. (2022)** Mazzaglia, P., Verbelen, T., Catal, O. & Dhoedt, B. The Free Energy Principle for Perception and Action: A Deep Learning Perspective. Entropy 24, 301, (2022).
311. **McFadden (2020)** McFadden, J. Integrating information in the brain's EM field: the cemi field theory of consciousness. Neurosci. Conscious. 2020, niaa016 (2020)

312. **McMillen & Levin (2024)** Patrick McMillen & Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. Communications Biology (2024) 7:378

313. **Medium (2024)** – Free publication platform, https://medium.com

314. **Melloni et al. (2021)** Melloni, L., Mudrik, L., Pitts, M., & Koch, C. Making the hard problem of consciousness easier. Science, 372(6545), 911-912. (2021, May 28).

315. **Merker (2007)** Merker, B. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. Behav. Brain Sci. 30, 63–81; discussion 81–134 (2007).

316. **Meta (2024)** The LCM team. Large Concept Models: Language Modeling in a Sentence Representation Space. FAIR at Meta. December 12, 2024

317. **Meta AI (2023a)** Meta AI. Introducing Llama 2. https://ai.meta.com/llama/ (2023)

318. **Meta AI (2023b)** Meta AI. I-JEPA: The first AI model based on Yann LeCun's vision for more human-like AI. https://ai.facebook.com/blog/yann-lecun-ai-model-i-jepa/ June 13, 2023

319. **Michaud et al. (2023)** Eric J. Michaud, Ziming Liu, Uzay Girit and Max Tegmark. The Quantization Model of Neural Scaling. arXiv: 2303.13506v1 [cs.LG] 23 Mar 2023

320. **Microsoft (2024)** Microsoft Research AI Frontiers (Team). Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. https://www.microsoft.com/en-us/research/articles/magentic-one-a-generalist-multi-agent-system-for-solving-complex-tasks/ November 4, 2024

321. **Miller et al. (2016)** Miller, A. H., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., and Weston, J. (2016). Key-value memory networks for directly reading documents. In EMNLP-16.

322. **Mineault (2023)** Patrick Mineault. The good old days of NeuroAI. How can you define a field as you're building it? https://naix.substack.com/p/the-good-old-days-of-neuroai (2023)

323. **Mineault et al. (2024)** Patrick Mineault, Niccolò Zanichelli, Joanne Zichen Peng, Anton Arkhipov, Eli Bingham, Julian Jara-Ettinger, Emily Mackevicius, Adam Marblestone, Marcelo Mattar, Andrew Payne, Sophia Sanborn, Karen Schroeder, Zenna Tavares, Andreas Tolias. NeuroAI for AI Safety. arXiv:2411.18526v1 [cs.AI] 27 Nov 2024

324. **Minsky (1986)** Minsky, M. L. The Society of Mind. Simon and Schuster, 1986

325. **Minsky (2007)** Minsky, M. L. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon and Schuster, 2007

326. **Mitchell & Krakauer (2022)** Melanie Mitchell and David C. Krakauer. The Debate Over Understanding in AI's Large Language Models. arXiv:2210.13966v2 [cs.LG] 27 Oct 2022

327. **Modis (2002)** Theodore Modis. Forecasting the growth of complexity and change. Technological Forecasting and Social Change, 69:377-404, 2002.

328. **Moravec (2003)** Hans Moravec. Simpler equations for vinge's technological singularity. http://www.frc.ri.cmu.edu/users/hpm/project.archive/robot.papers/2003/singularity2.html.

329. **Morris (1971)** Morris Ch. W. Foundations of the Theory of Signs. Signs and the Act / Morris Ch. Writings on the General Theory of Signs. Mouton and Co. Publishers, The Hague — Paris, 1971.

330. **Morris et al. (2023)** Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462 [cs.AI]

331. **Mumford (1991)** Mumford, D. On the computational architecture of the neocortex. Biological Cybernetics, 65, 135-145. (1991)

332. **Munkhdalai, Faruqui & Gopal (2024)** Tsendsuren Munkhdalai, Manaal Faruqui and Siddharth Gopal. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. arXiv:2404.07143v1 [cs.CL] 10 Apr 2024

333. **Nave et al. (2022)** Nave, K., Deane, G., Miller, M., & Clark, A. Expecting some action: Predictive processing and the construction of conscious experience. Review of Philosophy and Psychology, 13(4), 2022, 1019–1037.

334. **NeuroAI (2023)** NeuroAI paper browser. https://patrickmineault-neuroai-tree-scriptspaper-umap-ph4gak.streamlit.app/ (2023)
335. **Neurohive (2022)** https://neurohive.io/en/
336. **Newell (1990)** Newell, A. Unified Theories of Cognition (Harvard University Press, Cambridge, 1990).
337. **Newman (2010)** M. Newman, Networks: An introduction (Oxford University Press, 2010).
338. **NVIDIA (2024)** NVIDIA Quantum. Accelerating the future of scientific discovery. https://www.nvidia.com/en-us/solutions/quantum-computing/
339. **O'Regan & Noë (2001)** O'Regan, J. K. & Noë, A. A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. 24, 939–973; discussion 973–1031 (2001)
340. **Oizumi et al (2014)** Oizumi, M., Albantakis, L. & Tononi, G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. 10, e1003588 (2014)
341. **Olah & Carter (2017)** Chris Olah and Shan Carter. "Research Debt". In: Distill (2017).
342. **OpenAI (2018)** OpenAI Charter, 2018. URL https://openai.com/charter. Accessed October 12, 2023.
343. **OpenAI (2023a)** OpenAI. Gpt-4 technical report, 2023. arXiv: 2303.08774v2 [cs.CL] 16 Mar 2023
344. **OpenAI (2023b)** OpenAI. GPT-4 System Card. The Appendix of [OpenAI (2023a)]
345. **OpenAI (2023c)** OpenAI. Improving mathematical reasoning with process supervision. May 31, 2023 https://openai.com/research/improving-mathematical-reasoning-with-process-supervision.
346. **OpenAI (2023d)** OpenAI. Introducing Superalignment https://openai.com/blog/introducing-superalignment. July 5, 2023
347. **P´osfai et al. (2022)** M´arton P´osfai, Bal´azs Szegedy, Iva Baˇci´c, Luka Blagojevi´c, Mikl´os Ab´ert, J´anos Kert´esz, L´aszl´o Lov´asz, and Albert-L´aszl´o Baraba'si. Understanding the impact of physicality on network structure. arXiv:2211.13265v1 [cond-mat.stat-mech] 23 Nov 2022
348. **Pan et al. (2024)** Xudong Pan, Jiarun Dai, Yihe Fan, Min Yang. Frontier AI systems have surpassed the self-replicating red line. arXiv:2412.12140v1 [cs.CL] 9 Dec 2024
349. **Pang et al. (2023)** James C. Pang, Kevin M. Aquino, Marianne Oldehinkel, Peter A. Robinson, Ben D. Fulcher, Michael Breakspear, Alex Fornito. Geometric constraints on human brain function. Nature, volume 618, pages 566–574 (2023)
350. **Paolo, Gonzalez-Billandon & K´egl (2024)** Giuseppe Paolo, Jonas Gonzalez-Billandon, Bal ´azs K´egl. A Call for Embodied AI. arXiv:2402.03824v3 [cs.AI] 18 Jul 2024
351. **Parisi et al. (2019)** German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. Neural Networks, 113 (2019):54–71
352. **Park & Tallon-Baudry (2014)** Park, H. D. & Tallon-Baudry, C. The neural subjective frame: from bodily signals to perceptual consciousness. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369, 20130208 (2014)
353. **Park et al. (2024a)** Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, Hidenori Tanaka. Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space. arXiv:2406.19370v2 [cs.LG] 4 Nov 2024
354. **Park et al. (2024b)** Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, Michael S. Bernstein. Generative Agent Simulations of 1,000 People. arXiv:2411.10109v1 [cs.AI] 15 Nov 2024
355. **Parr et al. (2022)** Parr, T., Pezzulo, G. & Friston, K.J. Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. (MIT Press, 2022).

356. **Parvizi & Damasio (2001)** Parvizi, J. & Damasio, A. Consciousness and the brainstem. Cognition 79, 135–160 (2001)

357. **Pavlović et al. (2024)** Jelena Pavlović, Jugoslav Krstić, Luka Mitrović, Đorđe Babić, Adrijana Milosavljević, Milena Nikolić, Tijana Karaklić & Tijana Mitrović. Generative AI as a metacognitive agent: A comparative mixed-method study with human participants on ICF-mimicking exam performance. arXiv:2405.05285v1 [cs.HC] 7 May 2024

358. **Pazem et al. (2024)** Joséphine Pazem, Marius Krumm, Alexander Q. Vining, Lukas J. Fiderer, Hans J. Briegel. Free Energy Projective Simulation (FEPS): Active inference with interpretability. arXiv:2411.14991v1 [cs.AI] 22 Nov 2024

359. **PE (1996)** Concise Encyclopedia of Psychology. Second edition. Edited by Raymond J. Corsini, Alan J. Auerbach. John Wiley & Sons, Inc. 1996.

360. **Pearl & Mackenzie (2018)** Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect, 1st edn. Basic Books, Inc., New York (2018)

361. **Peirce (1931)** Charles Sanders Peirce. Collected Papers of Charles Sanders Peirce. Collected Papers of Charles Sanders Peirce v. 5. Harvard University Press, 1931.

362. **Peirce (1960)** Charles Sanders Peirce. Speculative grammar, ch. IV [Propositions]) / Papers of Charles Sanders Peirce. Vol. II: Elements of Logic. Edited by Charles Hartshorne and Paul Weiss. Harvard University Press, Cambridge (Mass.), 1960.

363. **Pellert et al. (2024)** Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspectives on Psychological Science 1–19. (2024)

364. **Pennartz (2018)** Pennartz, C. M. A. Consciousness, representation, action: the importance of being goal-directed. Trends Cogn. Sci. 22, 137–153 (2018)

365. **Perez (2023)** Carlos E. Perez. @IntuitMachine https://twitter.com/IntuitMachine (March 2024)

366. **Perez (2024)** Carlos E. Perez. @IntuitMachine https://x.com/IntuitMachine (June 2024)

367. **Pezzulo et al. (2024)** Giovanni Pezzulo, Thomas Parr, Paul Cisek, Andy Clark, and Karl Friston. Generating meaning: active inference and the scope and limits of passive AI. Trends in Cognitive Sciences, February 2024, Vol. 28, No. 2

368. **Pezzulo, Parr & Friston (2024)** Giovanni Pezzulo, Thomas Parr and Karl Friston. Active inference as a theory of sentient behavior. Biological Psychology 186 (2024) 108741

369. **Piaget (1979)** Jean Piaget. La psychogenèse des connaissances et sa signification épistémologique. Schèmes d'actionet apprentissage du langage / «Théorie du langage. Théorie de l'apprentissage». Le débat entre Jean Piaget et Noam Chomsky. Organisé et recueilli par Massimo Piatelli-Palmarini. Editions du Seuil. Paris, 1979

370. **Pichai & Hassabis (2023)** Sundar Pichai & Demis Hassabis. Introducing Gemini: our largest and most capable AI model. https://blog.google/technology/ai/google-gemini-ai/

371. **Pilarski et al. (2022)** Pilarski, P. M., Butcher, A., Davoodi, E., Johanson, M. B., Brenneis, D. J. A., Parker, A. S. R., Acker, L., Botvinick, M. M., Modayil, J., White, A. The Frost Hollow experiments: Pavlovian signalling as a path to coordination and communication between agents, arXiv:2203.09498. 2022

372. **Pinker (2003)** Steven Pinker. The Language Instinct: How the Mind Creates Language. Penguin UK, 2003.

373. **Porter et al. (2022)** Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. Predictors and consequences of intellectual humility. Nature Reviews Psychology, 1(9), 524–536. (2022)

374. **Premack (2004)** David Premack. Is language the key to human intelligence? Science, 303(5656):318–320, 2004.
375. **Prigogine & Stengers (1984)** Prigogine, Ilya; Stengers, Isabelle. Order out of Chaos: Man's new dialogue with nature. Flamingo, 1984.
376. **Prinz (2012)** Prinz, J. The Conscious Brain: How Attention Engenders Experience (Oxford Univ. Press, 2012)
377. **Priorelli & Stoianov (2024)** Matteo Priorelli & Ivilin Peev Stoianov. Deep Hybrid Models: Infer And Plan In The Real World. arXiv:2402.10088v2 [cs.RO] 21 Jun 2024
378. **Ramstead et al. (2021)** Maxwell J.D. Ramstead, Casper Hesp, Alexander Tschantz, Ryan Smith, Axel Constant, and Karl Friston. “Neural and phenotypic representation under the free-energy principle”. In: Neuroscience & Biobehavioral Reviews 120 (2021), pp. 109–122.
379. **Ramstead et al. (2022)** Maxwell J.D., Ramstead, Dalton A.R. Sakthivadivel, Conor Heins, Magnus T. Koudahl, Beren Millidge, Lancelot Da Costa, Brennan Klein, and Karl Friston. “On Bayesian mechanics: A physics of and by beliefs”. In: 10.48550/arXiv.2205.11543 (2022).
380. **Räuker et al. (2023)** Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. “Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks”. In: (Aug. 2023).
381. **Reardon (2019)** Reardon, S. Outlandish’ competition seeks the brain’s source of consciousness. Retrieved from sciencemag.org: doi: 10.1126/science.aaz8800 (2019, October 16).
382. **RM for BM (2022)** A Roadmap for Big Model. Produced by Beijing Academy of Artificial Intelligence (BAAI). 2022
383. **Robbins (2017)** Robbins Philip, "Modularity of Mind", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2017/entries/modularity-mind
384. **Rosenberg et al. (2024)** Louis Rosenberg, Gregg Willcox, Hans Schumann, Ganesh Mani. Towards Collective Superintelligence: Amplifying Group IQ using Conversational Swarms. arXiv:2401.15109v1 [cs.HC] 25 Jan 2024
385. **Rosenthal (2005)** Rosenthal, D. Consciousness and Mind (Clarendon, 2005)
386. **Rouleau & Levin (2025)** Nicolas Rouleau and Michael Levin. Brains and Where Else? Mapping Theories of Consciousness to Unconventional Embodiments. OSF PsyArXiv Preprints doi.org/10.31234/osf.io/va5mk Created: January 15, 2025 | Last edited: January 19, 2025
387. **Roy (2005)** Deb Roy. “Semiotic schemas: A framework for grounding language in action and perception”. In: Artificial Intelligence 167.1-2 (2005), pp. 170–205.
388. **RUCAIBox (2023)** RUCAIBox group (team). A Survey of Large Language Models. arXiv:2303.18223v13 [cs.CL] 24 Nov 2023
389. **Russell & Norvig (2021)** Stuart J. Russell and Peter Norvig. Artificial intelligence: a modern approach. Fourth edition. Pearson. 2021
390. **Russell (2019)** Stuart Russell. Human compatible: Artificial intelligence and the problem of control. Viking, 2019.
391. **Russell (2021)** Stuart Russell. Human-Compatible Artificial Intelligence. Computer Science Division, University of California, Berkeley. 2021
392. **Sablé-Meyer (2022)** Mathias Sablé-Meyer. Supervised by Stanislas Dehaene. Human Cognition of Geometric Shapes, a Window into the Mental Representation of Abstract Concepts. PhD Thesis. PSL/Collège de France. 2022
393. **Sandberg (2013)** Anders Sandberg. An overview of models of technological singularity. Future of Humanity Institute, Oxford University, 2013.
394. **Schacter et al. (2019)** Schacter DL, Gilbert DT, Nock MK, et al.. Psychology, 5th ed. New York, New York: Worth. 2019

395. **Schuld & Petruccione (2021)** Schuld, M. & Petruccione, F. Machine Learning with Quantum Computers (Springer, 2021)
396. **Schuurmans, Dai & Zanini (2024)** Dale Schuurmans, Hanjun Dai, Francesco Zanini. Autoregressive Large Language Models are Computationally Universal. arXiv:2410.03170v1 [cs.CL] 4 Oct 2024
397. **Sequoia Cap. (2022)** Sequoia Capital. Generative AI: A Creative New World. url: https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/
398. **Sequoia Cap. (2023)** Sequoia Capital. The New Language Model Stack. How companies are bringing AI applications to life. https://www.sequoiacap.com/article/llm-stack-perspective/
399. **Seth & Bayne (2022)** Anil K. Seth and Tim Bayne. Theories of consciousness. Nature Reviews. Neuroscience. Vol. 23. 429-452 (July 2022)
400. **Seth & Tsakiris (2018)** Seth, A. K. & Tsakiris, M. Being a beast machine: the somatic basis of selfhood. Trends Cogn. Sci. 22, 969–981 (2018)
401. **Seth (2015)** Seth, A. K. in Open MIND (eds Windt, J. M. & Metzinger, T.) (MIND Group, 2015)
402. **Seth (2021)** Seth, A. K. Being You: A New Science of Consciousness (Faber & Faber, 2021)
403. **Shai (2024)** Adam Shai. Transformers Represent Belief State Geometry in their Residual Stream. https://www.alignmentforum.org/posts/gTZ2SxesbHckJ3CkF 17th Apr 2024
404. **Shanahan (2015)** Murray Shanahan. The Technological Singularity. MIT Press, August 2015.
405. **Shanahan (2024)** Murray Shanahan. Simulacra as Conscious Exotica. arXiv:2402.12422v1 [cs.AI] 19 Feb 2024
406. **Shanahan et al. (2023)** Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-Play with Large Language Models. arXiv:2305.16367v1 [cs.CL] 25 May 2023
407. **Sharma (2024)** Ankit Sharma. Bridging Paradigms: The Integration of Symbolic and Connectionist AI in LLM Driven Autonomous Agents. Journal of Artificial Intelligence General Science. V. 6 , 2024
408. **Shevlane et al. (2023)** Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano and Allan Dafoe. Model evaluation for extreme risks. DeepMind. arXiv:2305.15324v1 [cs.AI] 24 May 2023
409. **Shulman & Bostrom (2021)** Shulman, C., & Bostrom, N. Sharing the world with digital minds. S. Clarke, H. Zohny, & J. Savulescu (Eds.), Rethinking Moral Status. Oxford University Press, 2021
410. **SI (2022)** http://singinst.org/overview/whatisthesingularity - Singularity Institute 2022
411. **Simon (2017)** Simon, J. A. Vagueness and zombies: Why ‘phenomenally conscious’ has no borderline cases. Philosophical Studies, 174(8), 2017, 2105–2123.
412. **Sloman (2021)** Steven A. Sloman et al, Cognitive Neuroscience Meets the Community of Knowledge, Frontiers in Systems Neuroscience, 2021
413. **Smirnov, Ponomarev and Agafonov (2024)** Alexander Smirnov, Andrew Ponomarev and Anton Agafonov. Ontology-Based Neuro-Symbolic AI: Effects on Prediction Quality and Explainability. Article in IEEE Access · October 2024
414. **Snooks (2005)** Snooks G.D. Why is history getting faster? Measurement and explanation // Философские науки 2005, N4, с.51-69
415. **Solms (2018)** Solms, M. The hard problem of consciousness and the free energy principle. Front. Psychol. 9, 2714 (2018)
416. **Solms (2021)** Solms, M. The Hidden Spring: A Journey to the Source of Consciousness (Profile Books, 2021).
417. **Solomonoff (1985)** Ray J. Solomonoff. The time scale of artificial intelligence: reflections on social effects. Nort-Holland Human Systems Management, 5:149-153, 1985.

418. **Sorensen et al. (2024)** Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, Yejin Choi. A Roadmap to Pluralistic Alignment. arXiv:2402.05070v1 [cs.AI] 7 Feb 2024
419. **Sotala & Yampolskiy (2016)** Kaj Sotala and Roman V Yampolskiy. Responses to catastrophic AGI risk: a survey. Physica Scripta, 90(1):018001, 2014.
420. **Sporns et al. (2004)** Sporns, O., Chialvo, D.R., Kaiser, M., and Hilgetag, C.C. Organization, development and function of complex brain networks. Trends Cogn. Sci. 8, 418–425. (2004)
421. **State of AI (2024)** - State of AI Bi-Weekly Roundups (Summary of Frontier AI Research) https://stateai.substack.com/
422. **Sternberg (1998)** Sternberg, R. J. A balance theory of wisdom. Review of General Psychology, 2, 347–365. (1998)
423. **Street et al. (2024)** Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael Mckibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, Robin I. M. Dunbar. LLMs achieve adult human performance on higher-order theory of mind tasks. arXiv:2405.18870v2 [cs.AI] 31 May 2024
424. **Substack (2024)** – Free publication platform, https://substack.com
425. **Suddendorf & Corballis (2007)** Thomas Suddendorf and Michael C Corballis. The evolution of foresight: What is mental time travel, and is it unique to humans? Behavioral and brain sciences, 30(3):299–313, 2007.
426. **Suddendorf et al. (2011)** Thomas Suddendorf, Donna Rose Addis, and Michael C Corballis. Mental time travel and shaping of the human mind. M. Bar, pp. 344–354, 2011.
427. **Suleyman & Bhaskar (2023)** Mustafa Suleyman and Michael Bhaskar. The Coming Wave: Technology, Power, and the 21st Century's Greatest Dilemma. Crown, September 2023.
428. **Sumers et al. (2024)** Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, Thomas L. Griffiths. Cognitive Architectures for Language Agents. arXiv:2309.02427v3 [cs.AI] 15 Mar 2024
429. **Sun et al. (2024)** Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, Zhenguo Li. A Survey of Reasoning with Foundation Models. arXiv:2312.11562v5 [cs.AI] 25 Jan 2024
430. **SuperAGI (2024)** SuperAGI team. Veagle: Advancements in Multimodal Representation Learning. arXiv:2403.08773v1 [cs.CV] 18 Jan 2024
431. **Sutton & Barto (2018)** Sutton, R. S., Barto, A. G. Reinforcement Learning: An Introduction, second edition. MIT Press, 2018
432. **Sutton (2016)** Sutton, R. The Future of Artificial Intelligence Belongs to Search and Learning, Talk at the University of Toronto, https://youtu.be/fztxE3Ga8kU. 2016
433. **Sutton (2019)** Sutton, R. The bitter lesson. Incomplete Ideas (blog), http://www.incompleteideas.net/IncIdeas/BitterLesson.html. 2019
434. **Sutton (2022)** Sutton, R. S. The quest for a common model of the intelligent decision maker. In: Multi-disciplinary Conference on Reinforcement Learning and Decision Making, arXiv:2202.13252, 2022
435. **Sutton et al. (2011)** Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., Precup, D. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In Proceedings of the 10th International Conference on Autonomous Agents and Multi-agent Systems, 2011, Volume 2, pp. 761-768.

436. **Sutton et al. (2022)** Sutton, R. S., Machado, M. C., Holland, G. Z., Timbers, D. S. F., Tanner, B., White, A. Reward-respecting subtasks for model-based reinforcement learning, arXiv:2202.03466, 2022
437. **Sutton et al. (2023)** Richard S. Sutton, Michael Bowling, and Patrick M. Pilarski. The Alberta Plan for AI Research. arXiv:2208.11173v3 [cs.AI] 21 Mar 2023
438. **Taagepera (1979)** R Taagepera. People, skills, and resources: an interaction model for world population growth. Technological forecasting and social change, 13:13-30, 1979.
439. **Tao et al. (2024)** Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, Jingren Zhou. A Survey on Self-Evolution of Large Language Models. arXiv:2404.14387v1 [cs.CL] 22 Apr 2024
440. **Tay et al. (2023)** Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Bahri, D., Schuster, T., Zheng, S., Zhou, D., Houlsby, N., and Metzler, D. UL2: Unifying language learning paradigms. In The Eleventh International Conference on Learning Representations, 2023.
441. **Tegmark (2018)** Tegmark, M. Life 3.0: Being human in the age of artificial intelligence (Vintage, 2018).
442. **Thomas (2024)** Rosemary J Thomas, The Rise of Large Action Models, LAMs: How AI Can Understand and Execute Human Intentions? Published in Version 1, Jan 16, 2024
443. **Tian et al. (2024)** Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, Dong Yu. Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing. arXiv:2404.12253v1 [cs.CL] 18 Apr 2024
444. **Together AI (2024)** Together AI. Mixture-of-Agents Enhances Large Language Model Capabilities. arXiv:2406.04692v1 [cs.CL] 7 Jun 2024
445. **Tononi & Edelman (1998)** Tononi, G. & Edelman, G. M. Consciousness and complexity. Science 282, 1846–1851 (1998)
446. **Tononi & Koch (2015)** Tononi, G., & Koch, C. Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society of London B: Biological Sciences, 370 (1668). (2015)
447. **Tononi (2004)** Tononi, G. An information integration theory of consciousness. BMC Neuroscience 5, 42-72. (2004)
448. **Tononi (2008)** Tononi, G. Consciousness as integrated information: a provisional manifesto. Biol. Bull. 215, 216–242 (2008).
449. **Tononi (2012)** Tononi, G. Integrated information theory of consciousness: an updated account. Arch. Ital. Biol. 150, 293–329 (2012).
450. **Tononi et al. (2016)** Tononi, G., Boly, M., Massimini, M. & Koch, C. Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. 17, 450–461 (2016).
451. **Tulving (2002)** Endel Tulving. Episodic memory: From mind to brain. Annual review of psychology, 53(1):1–25, 2002.
452. **Turchin (1977)** Valentin Turchin. The Phenomenon of Science. A cybernetic approach to human evolution. Columbia University Press, New York, 1977.
453. **Turing (1937)** Turing, A. M. On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 2, 230-265. (1937)
454. **Turing (1945)** Turing, A. M. Proposal for development in the Mathematics Division of an Automatic Computing Engine (ACE). Report E.882, Executive Committee, NPL, Mathematics. (1945)
455. **Turing (1950)** A.M. Turing. Computing Machinery and Intelligence. Mind, LIX:433–460, October 1950.
456. **Van Gigch (1974)** John P. van Gigch. Applied general systems theory. Harper & Row PublishersL, 1974
457. **Van Mieghem (2010)** Van Mieghem, Graph spectra for complex networks (Cambridge University Press, 2010).

458. **Vaswanj et al. (2023)** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 2023. arXiv.1706.03762
459. **Vilas et al. (2024)** Martina G. Vilas, Federico Adolfi, David Poeppel, Gemma Roig. Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience. arXiv:2406.01352v1 [cs.AI] 3 Jun 2024
460. **Vinge (1993)** Vernor Vinge. The coming technological singularity: How to survive in the posthuman era. Number NASA CP-10129, 1993.
461. **Volzhenin et al. (2022)** Konstantin Volzhenin, Jean-Pierre Changeux and Guillaume Dumasa, Multilevel development of cognitive abilities in an artificial neural network. PNAS 2022 Vol. 119 No. 39 e2201304119
462. **Von Neumann & Burks (1966)** Von Neumann, J. & Burks, A. W. Theory of Self Reproducing Automata (University of Illinois Press, 1966).
463. **Waade et al. (2024)** Peter Thestrup Waade, Christoffer Lundbak Olesen, Jonathan Ehrenreich Laursen, Samuel William Nehrer, Conor Heins, Karl Friston and Christoph Mathys. As One and Many: Relating Individual and Emergent Group-Level Generative Models in Active Inference. Preprints.org. doi:10.20944/preprints202410.1895.v1 24 October 2024
464. **Wang et al. (2023)** Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, Explain, Plan and Select: Interactive planning with large language models enables open-world multi-task agents. arXiv preprint arXiv:2302.01560, 2023.
465. **Watson & Levin (2023)** Richard Watson and Michael Levin. The collective intelligence of evolution and development. Collective Intelligence Volume 2:2: 1–22 © The Author(s) 2023
466. **Webb & Graziano (2015)** Webb, T. W., & Graziano, M. S. A. The attention schema theory: A mechanistic account of subjective awareness. Frontiers in Psychology, 6. 2015
467. **Webb et al. (2023)** Taylor Webb, Keith J. Holyoak, Hongjing Lu. Emergent Analogical Reasoning in Large Language Models. arXiv:2212.09196v3 [cs.AI] 3 Aug 2023
468. **Wendler et al. (2024)** Chris Wendler, Veniamin Veselovsky, Giovanni Monea, Robert West. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. arXiv:2402.10588v3 [cs.CL] 5 Jun 2024
469. **Weng (2018)** Lilian Weng. Meta-learning: Learning to learn fast. lilianweng.github.io/lil-log, 2018.
470. **West et al. (2010)** D. B. West et al., Introduction to graph theory, (Prentice hall Upper Saddle River, 2001).
471. **Weston & Sukhbaatar (2023)** Jason Weston and Sainbayar Sukhbaatar. System 2 Attention (is something you might need too). arXiv:2311.11829v1 [cs.CL] 20 Nov 2023
472. **Weyns (2020)** Danny Weyns. An Introduction to Self-adaptive Systems : A Contemporary Software Engineering Perspective. Wiley-IEEE Computer Society Pr. 2020
473. **Weyns et al. (2022)** Danny Weyns, Thomasb Back, Rene Vidal, Xin Yao, and Ahmed Nabile Belbachir. The vision of self-evolving computing systems. Journal of Integrated Design and Process Science 26, 3-4 (2022), 351–367.
474. **Whyte (2019)** Whyte, C. J. Integrating the global neuronal workspace into the framework of predictive processing: Towards a working hypothesis. Consciousness and Cognition, 73, 102763. 2019
475. **Wierzbicka (1972)** Anna Wierzbicka. Semantic Primitives. Introduction / WierzbickaA. Semantic Primitives. Frankfurt-a / M., 1972.

476. **Wilson (2024)** Jonathan Jared Wilson. Quantifying Consciousness in Artificial Intelligence: An Integrated Approach Using Quantum Mechanics, Information Theory, and Neuroscience (March 24, 2024). Available at SSRN: https://ssrn.com/abstract=4770970

477. **Wolf et al. (2023)** Yotam Wolf, Noam Wies, Yoav Levine, Amnon Shashua. Fundamental limitations of alignment in large language models. arXiv:2304.11082v1 [cs.CL] 19 Apr 2023

478. **Wolpert & Kinney (2024)** David H. Wolpert and David B. Kinney. A Stochastic Model of Mathematics and Science. arXiv:2209.00543v2 [math.LO] 14 Mar 2023

479. **Wolpert (2022)** David H. Wolpert. What can we know about that which we cannot even imagine? arXiv:2208.03886v1 [physics.hist-ph] 8 Aug 2022

480. **Wonga et al. (2023)** Michael L. Wonga, Carol E. Clelandc, Daniel Arend Jr., Stuart Bartlettd, H. James Cleaves IIa, Heather Demarestc, Anirudh Prabhua, Jonathan I. Lunineg, Robert M. Hazena. On the roles of function and selection in evolving systems. PNAS 2023 Vol. 120 No. 43 e2310223120

481. **Wozniak (2010)** Steve Wozniak. Could a Computer Make a Cup of Coffee? Fast Company interview: https://www.youtube.com/watch?v=MowergwQR5Y, 2010.

482. **Wu et al. (2024)** Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, Yoon Kim. The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities. arXiv:2411.04986v1 [cs.CL] 7 Nov 2024

483. **xAI (2023)** xAI Team. Announcing Grok. https://x.ai/, November 4, 2023

484. **Xiong et al. (2024a)** Zheyang Xiong, Ziyang Cai, John Cooper, Albert Ge, Vasilis Papageorgiou, Zack Sifakis, Angeliki Giannou, Ziqian Lin, Liu Yang, Saurabh Agarwal, Grigorios G Chrysos, Samet Oymak, Kangwook Lee, Dimitris Papailiopoulos. Everything Everywhere All at Once: LLMs can In-Context Learn Multiple Tasks in Superposition. arXiv:2410.05603v1 [cs.LG] 8 Oct 2024

485. **Xiong et al. (2024b)** Haoyi Xiong, Zhiyuan Wang, Xuhong Li, Jiang Bian, Zeke Xie, Shahid Mumtaz, Laura E. Barnes. Converging Paradigms: The Synergy of Symbolic and Connectionist AI in LLM-Empowered Autonomous Agents. arXiv:2407.08516v1 [cs.AI] 11 Jul 2024

486. **Yamakawa (2024)** Hiroshi Yamakawa. Investigating Alternative Futures: Human and Superintelligence Interaction Scenarios https://www.lesswrong.com/posts/QeqKjDQM7WaKsSCMt 4th Jan 2024

487. **Yang et al. (2024)** Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, Bin Cui. Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models. arXiv:2406.04271v1 [cs.CL] 6 Jun 2024

488. **Yaron et al. (2022)** Itay Yaron, Lucia Melloni, Michael Pitts and Liad Mudrik. The Consciousness Theories Studies (ConTraSt) database for analyzing and comparing empirical studies of consciousness theories. Nature Human Behaviour, 6(4), 2022, pp.593–604.

489. **Yin et al. (2024)** Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Tianxiang Sun, Cheng Chang, Qinyuan Cheng, Ding Wang, Xiaofeng Mou, Xipeng Qiu, Xuanjing Huang. Aggregation of Reasoning: A Hierarchical Framework for Enhancing Answer Selection in Large Language Models. arXiv:2405.12939v1 [cs.CL] 21 May 2024

490. **Yuan W. et al. (2024)** Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, Jason Weston. Self-Rewarding Language Models. arXiv:2401.10020v1 [cs.CL] 18 Jan 2024

491. **Yuan Z. et al. (2024)** Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, Lichao Sun. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework. arXiv:2403.13248v2 [cs.CV] 22 Mar 2024

492. **Yudkowsky (2007)** Eliezer S. Yudkowsky. Three major singularity schools. http://yudkowsky.net/singularity/schools, 2007.

493. **Yudkowsky et al. (2008)** Eliezer Yudkowsky et al. Artificial Intelligence as a positive and negative factor in global risk. Global catastrophic risks, 1(303):184, 2008.

494. **Yuksekgonul et al. (2024)** Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, James Zou. TEXTGRAD: Automatic "Differentiation" via Text. arXiv:2406.07496v1 [cs.CL] 11 Jun 2024

495. **Zador et al. (2022)** Anthony Zador, Blake Richards, Bence Ölveczky, Sean Escola, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Koerding, Alexei Koulakov, et al. "Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution". In: (Oct. 2022).

496. **Zelikman et al. (2024)** Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, Noah D. Goodman. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. arXiv:2403.09629v2 [cs.CL] 18 Mar 2024

497. **Zhang & Xu (2024)** Zhengquan Zhang and Feng Xu. An Overview of the Free Energy Principle and Related Research. Neural Computation 36, 963–1021 (2024)

498. **Zhang E. et al. (2024)** Edwin Zhang, Vincent Zhu, Anat Kleiman, Naomi Saphra, Benjamin L. Edelman, Milind Tambe, Sham Kakade, Eran Malach. Transcendence: Generative Models Can Outperform The Experts That Train Them. arXiv:2406.11741v1 [cs.LG] 17 Jun 2024

499. **Zhang et al. (2024)** Shiyang Zhang, Aakash Patel, Syed A Rizvi, Nianchen Liu, Sizhuang He, Amin Karbasi, Emanuele Zappala, David van Dijk. Intelligence at the Edge of Chaos. arXiv:2410.02536v2 [cs.AI] 8 Oct 2024

500. **Zhang J. et al. (2024)** Jianyu Zhang, Niklas Nolte, Ranajoy Sadhukhan, Beidi Chen, L´eon Bottou. Memory Mosaics. arXiv:2405.06394v2 [cs.LG] 13 May 2024

501. **Zhao et al. (2021)** Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, and Yoshua Bengio. A consciousnessinspired planning agent for model-based reinforcement learning. Advances in Neural Information Processing Systems, 34, 2021.

502. **Zhao Feifei et al. (2024)** Feifei Zhao, Hui Feng, Haibo Tong, Zhengqiang Han, Enmeng Lu, Yinqian Sun, Yi Zeng. Building Altruistic and Moral AI Agent with Brain-inspired Affective Empathy Mechanisms. arXiv:2410.21882v1 [cs.AI] 29 Oct 2024

503. **Zhao Siyun et al. (2024)** Siyun Zhao, Yuqing Yang, ZilongWang, Zhiyuan He, Luna K. Qiu, Lili Qiu. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv:2409.14924v1 [cs.CL] 23 Sep 2024

504. **Zhaozhiming (2024)** - Zhaozhiming. Advanced RAG Retrieval Strategies: Flow and Modular. https://ai.gopubby.com/advanced-rag-retrieval-strategies-flow-and-modular-672493acb4a7 Jun 12, 2024

505. **Zheng J. et al. (2024)** Junhao Zheng, Shengjie Qiu, Chengming Shi, Qianli Ma. Towards Lifelong Learning of Large Language Models: A Survey. arXiv:2406.06391v1 [cs.LG] 10 Jun 2024

506. **Zheng Y. et al. (2024)** Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Yongqiang Ma. LLAMAFACTORY: Unified Efficient Fine-Tuning of 100+ Language Models. arXiv:2403.13372v2 [cs.CL] 21 Mar 2024

507. **Zhou et al. (2024)** Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, Huaixiu Steven Zheng. SELF-DISCOVER: Large Language Models Self-Compose Reasoning Structures. arXiv:2402.03620v1 [cs.AI] 6 Feb 2024

508. **Zou et al. (2023)** Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, et al. "Representation engineering: A top-down approach to AI transparency". In: arXiv [cs.LG] (Oct. 2023).

509. **Альтшуллер (1979)** Альтшуллер Г. С. Творчество как точная наука. — М.: Советское радио, 1979
510. **Альтшуллер (2010)** Альтшуллер Г. С. Найти идею: Введение в ТРИЗ — теорию решения изобретательских задач, 3-е изд. — М.: Альпина Паблишер, 2010
511. **Буданов (2015**) Буданов В.Г. Синергетика и теория сложности: междисциплинарный подход. Часть I. Принципы. Методология. Курск: ЗАО «Университетская книга», 2015
512. **Волкова и Денисов (2001)** Волкова В. Н., Денисов А. А. Основы теории систем и системного анализа. - СПб.: СПбГТУ, 2001
513. **Карелов (2022)** https://sergey-57776.medium.com/ - Статьи Сергея Карелова (2022-2024)
514. **Назаретян (2017)** Назаретян А. П. "Нелинейное будущее", Изд. 4-е, М.:АРГАМАК-МЕДИА, 2017
515. **Новиков (2012)** Новиков А. Е**.** Система управления стратегическим развитием многопрофильной компании. – Saarbrücken.: LAP LAMBERT Academic Publishing, 2012
516. **Новиков (2017)** Новиков А. Е. Собрание сочинений. Том II. Бизнес 2001-2016. На правах рукописи. СПб. 2017
517. **Новиков (2022)** Новиков А. Е. Собрание сочинений. Том I. Разное 2000-2022. На правах рукописи. СПб. 2022
518. **Новиков (2023)** Новиков А. Е. СКАЙНЕТ 2022. Концепция Проекта Создания Сильного Искусственного Интеллекта. Системный Подход. На правах рукописи. СПб. 2023
519. **Панов (2014)** Панов А.Д. Технологическая сингулярность, теорема Пенроуза об искусственном интеллекте и квантовая природа сознания // Приложение к журналу "Информационные технологии", 2014, N5.
520. **Степанов (2001)** Семиотика: Антология/Сост. Ю. С. Степанов. Изд. 2-е, испр. и доп. – М.: Академический Проект; Екатеринбург: Деловая книга, 2001
521. **Фёдоров (1906)** Фёдоров Н. Ф. Философия общего дела. Т. 1. Верный, 1906
522. **Фёдоров (1913)** Фёдоров Н. Ф. Философия общего дела. Т. 2. М., 1913
523. **Харламов (2014)** Харламов А. А. Модель мира человека – семантическая сеть. Институт высшей нервной деятельности и нейрофизиологии РАН, Москва, 2021

Abbreviations

- 3D – Three-Dimensional
- AGI – Artificial General Intelligence
- AI – Artificial Intelligence
- AMI - Autonomous Machine Intelligence
- API - Application Programming Interface
- AR – Augmented Reality
- ASI – Artificial Super Intelligence
- ASC - Air Street Capital
- AST - Attention Schema Theory
- BAAI - Beijing Academy of Artificial Intelligence
- BC/AC – Before Christ / After Christ
- BM - Big Model
- CEO – Chief Executive Officer
- CF – Cash Flow
- CM – Conceptual Model
- CMI – International Management Center (Geneva)
- CNAS - Center for a New American Security
- CPU – Central Processing Unit
- CS – Control System
- CSP - Constraint Satisfaction Problems
- CTM - Conscious Turing Machine
- CoT – Chain of Thought
- CV – Computer Vision
- DB – DataBase
- DBMS – DataBase Management System
- DL – Deep Learning
- DS&S – Design Statement and Specification
- EBM - Energy-Based Model (Method)
- EM - ElectroMagnetic
- EU – European Union
- ExpMax - Expectation–Maximization
- FE – First Edition
- FS&ED- Feasibility Study and Exploratory Design
- GenAI – Generative AI
- GNWT - Global Neural Workspace Theory
- GPT – Generative Pretrained Transformer
- GPU – Graphic Processing Unit
- GR – Government Relations
- GST – General Systems Theory
- GWT - Global Workspace Theory
- GVF- Generalized Value Function
- HITL - Human-In-The-Loop

- HOT - Higher-Order Theory
- HRM – Human Resource Management
- IA – Intellectual Action
- IEMI – European Institute of International Management (Paris)
- IGT - Information Generation Theory
- IIT - Integrated Information Theory
- IQ - Intelligence Quotient
- IR – Investors Relations
- IS – Internal Space
- IT – Information Technology
- JEA - Joint Embedding Architecture
- JEPA - Joint Embedding Predictive Architecture
- KAN - Kolmogorov–Arnold Network
- KB – Knowledge Base
- KG - Knowledge Graph
- KPI – Key Performance Indicator
- KVM - Ray Kurzweil, Vernor Vinge, and Hans Moravec
- LAM – Large Action Model
- LCM – Large Concept Model
- LLM – Large Languages Model
- LSICS - Large Scale Intelligent Computing System
- LTM - Long Term Memory
- MACI - Multi-Agent Collaborative Intelligence
- MAD - ML/AI/Data
- MAS – Multi-Agent System
- MBRL – Model Based Reinforcement Learning
- MDP - Markov Decision Process
- ML – Machine Learning
- MM – Mental Map (Multi Modal)
- MoA – Mixture of Agents
- MoE – Mixture of Experts
- MV - MetaVers
- MTT – Mental Time Travel
- NAE – Novikov Alexander E. (Author)
- NLP – Natural Language Processing
- NMD – Normative Methodical Document
- NPU – Neural Processing Unit
- OS – Operating System
- PEAS – Performance, Environment, Actuators, Sensors
- PESTEL – Political, Economic, Social, Technological, Environment, Legal - (Analysis)
- POMDP – Partially Observable Markov Decision Process
- PPR&D – Pre-Project Research and Development
- PR – Public Relations
- PSS - Project Scope Statement
- QA – Question and Answer

- RAG - Retrieval-Augmented Generation
- RAM - Random-Access Memory (Operational)
- R&D – Research and Development
- RL – Reinforcement Learning
- RLAIF - Reinforcement Learning from AI Feedback
- RLHF - Reinforcement Learning from Human Feedback
- RNA - RiboNucleic Acid
- RPT - Recurrent Processing Theory
- S2A - System 2 Attention
- SA – System Analysis
- SAS – Self Adaptive System
- SE – Second Edition
- SI – Strong Intelligence
- SO – Self-Organization
- SOTA - State Of The Art
- SQL - Structured Query Language
- SSL - Self-Supervised Learning
- STEM - Science, Technology, Engineering, Mathematics
- STM - Short Term Memory
- STOMP - SubTask, Option, Model, Planning
- STP – Science Technical Progress
- SWOT - Strengths, Weaknesses, Opportunities, Threats - (Analysis)
- T2I – Text to Image
- T2V – Text to Video
- TAP – Theory of the Adjacent Possibly
- TCS - Theoretical Computer Science
- T&M – Theory and Metodology
- TOR – Terms of Reference
- TPU – Tensor Processing Unit
- UK – United Kingdom
- USA – United States of America
- USSR – Union of Soviet Socialistic Republics
- VICReg - Variance-Invariance-Covariance Regularization
- VR – Virtual Reality
- ТРИЗ - Theory of Inventive Problem Solving

Index