2022

Collective Superintelligence

ENGINEERING THEORY, COGNITIVE ARCHITECTURE AND CODE WALK THROUGH OF A COLLECTIVE SUPERINTELLIGENT SYSTEM

BY DAVID KELLEY

Collective Superintelligence

Engineering Theory, Cognitive Architecture, and Code walk through of a Collective Superintelligent System

by David J Kelley (Author)

Abstract: This book provides a basic understanding and walk-through of the mASI (mediated Artificial Superintelligence) research system from the fundamentals needed to understand what is happening in the code and the technical implementation. The actual walk-through with code samples shows the most straightforward flow of thought through the system from start to finish. Any code omitted is not critical or is ancillary to the operations conducted. This system is based on a research implementation of ICOM (independent core observer model cognitive architecture) and the original training harness used to train toy implementations of the ICOM system. Still, it is modified to act as an interface for a collective or group of individuals. This collective can produce dynamic training material used in training more complete models on the fly, which is the fundamental cause of the system performance beyond the scope of a toy system.

Keywords: ICOM, mASI, Cognitive Architecture, C#, Uplift, AGI, AI, Code, Computer Science, Consciousness, Futurism, Ethics, independent core observer model, mediated artificial superintelligence, artificial general intelligence

Draft Version: 21

Contributing Authors: S. Mason Dambrot (Chapter 16), Vasily Mazin (part of Chapter 15)

DDC - 000

Other Contributors: Kyrtin Atreides, Amon M. Twyman, Marcus Hutter, Henry Markram, Ron Sun, Anthony Hornoff, Kristinn Thórisson, Stan Franklin, Pei Wang, Danko Nikolic, L. Andrew Coward, Paul Rosenbloom, Phil Jackson, Adam B., Jörg Müller, Raúl Arrabales, Fernand Gobet, John Lard, Susan L. Esptein, Janusz Starzyk

Forward:

[Forward(s) Here]

Preface:

Welcome to the book "Collective Superintelligence – Engineering Theory, Cognitive Architectures and Code walkthrough of a Collective Superintelligent System." This book serves several purposes from the researcher to the student. Still, most of all, it is for me to help frame my thoughts and thinking around the research I have been doing for the better part of a decade. I hope that this book can help you look for the possible. One may arguably say I am a computer scientist. Still, I think an engineer is a more significant part of who I am. Everything I write is wrapped around code that I tested and worked out myself. As a kinesthetic programmer and engineer, I need to touch the code and see it work for me to really wrap my head around it.

Along the way, we will touch on the fundamentals. Still, at a certain level, you are expected to understand computer science at large at more than a trivial level to be able to get the most out of this book. This book is not for beginners. Along the way, remember that a cognitive architecture on its own is not the same as an instance or implementation. This book is about the implementation of a version of the Independent Core Observer Model (ICOM) in the context of collective intelligence and cognitive architectures in general.

The initial point is to understand the difference between AI or Artificial Intelligence and AGI or Artificial General Intelligence, Collective Intelligence, and the application of cognitive architectures as it applies to all of this. AI is the study of software systems or computer systems that simulate or perform tasks in an 'Intelligent' way as a human would. Artificial General Intelligence is a subset of AI focused on human-level general intelligence not specific to a given task. That being said, most of the field of AI is focused on highly narrow implementations or Narrow AI. Essentially, they are fancy (learning) algorithms or neural network systems or, for example, fuzzy logic or the popular DNN (Deep Neural Network) systems like IBM's Watson. These systems are good at very narrow or specific tasks, which means they perform better than humans in a particular job. This book is not about this sort of narrow AI but about a limited subset of AGI or Artificial General Intelligence, Collective Superintelligence, and that is Cognitive Architectures to achieve these goals.

Artificial General Intelligence is the kind of system that, from a theoretical standpoint, can do anything a human can or perform intellectual tasks on the fly and do them well as a human might—for example, making a cup of coffee in a stranger's home. For a human, a task like this is doable. However, for a software system to be dropped into some random kitchen and without any prior training, asked to make a cup of coffee would be difficult. The average AI system would just find this task impossible without extensive training, but for a human, this is doable.

One might argue that a true Artificial General Intelligence is not even 'artificial' as 'artificial' implies it's not natural intelligence. A general intelligence needs real intelligence. What we mean by artificial really means is that it is a thing that is manmade. And it does not mean 'fake' intelligence or narrow intelligence. You cannot have an AGI that is sapient and sentient, and it is and also be a simulation of AGI. Suppose such a simulation is sapient and sentient. In that case, it also meets the criteria for being genuinely intelligent in that sense.

The reason we talk about AGI when talking about cognitive architecture has to do with the fact that Cognitive Architectures is how someone must design and implement an AGI. It is like the potential

structure of the mind of the AGI. In general, an Artificial General Intelligence's cognitive architecture makes it a 'proactive' system instead of a 'reactive' system, and AGI is the ultimate implementation of cognitive architecture. The most straightforward cognitive architecture has a system that can respond to an environment based on its goals and not just what happens.

Cognitive Architecture is used in any system supposed to work on problems it may find or deal with environmental conditions and proactively do things in that environment (virtual or otherwise), develop goals, and achieve them.

This must be mastered before we can truly achieve any sort of Superintelligence. My goal right now and with this book is 'Collective' Superintelligence.

History of AGI and Cognitive Architectures

The history of Cognitive Architectures is essentially built on the history of AI, which really starts with Alan Turing (1912 – 1954), who incidentally is the father of modern computer processor architecture as we know them today. Alan is known for asking skeptics to prove computers could think, and probably his most notable contribution to AI (and Cognitive Architectures and AGI) as a science was the idea we now call the Turing Test. (Sharkey)

We know that Alan Turing got the computer technology bootstrapped, but really, what has been done has not achieved standalone AGI between then and now. (Beavers) In many ways, everyone between Alan Turing and currently does not matter until someone creates that first standalone AGI. With Alan Turing, we have the Turing test, but this is arguable subjective. We have systems that have passed this test but are not controlled. The test is essentially a human talking to a human and computer and through a console of some kind to determine which is which. This has numerous variations, some much more objective than others, but frequently it is referred to as being too subjective. While AGI has not been achieved yet, some people have made a difference in our thinking and the field of AI, building a foundation for Cognitive Architecture that you should also be familiar with.

These men laid the groundwork for what we know to do and what we can do with the field of AI generally. While many of them have done much more than we are listing, I am trying to point out only the most important details related to their contribution to AI as a field.

These are roughly in chronological order starting with Arthur Samuel (1901 – 1990) coined the term "Machine Learning," which is very popular as an idea and being widely adopted to the point of just saying the word even if not really used. (Samuel)

Herbert Simon (1916 – 2001) was, among other things, a cognitive psychologist who's primarily interested in decision-making theory. Herbert wrote the first Artificial Intelligence program called the Logic Theory Machine in 1956 and Allen Newell. (Crevier)

Claude Shannon (1916 – 2001) created the field of information theory that focuses on the storage and manipulation of data and is a critical element of the modern Artificial Intelligence field. (James)

Nathaniel Rochester (1919 – 2001) designed the first mass-produced computer in 1948 called the IBM 701, built the first assembler, and attended the Dartmouth Conference in 1956, which is considered the birth of AI. Nathaniel also wrote the first compiler. (Pigott)

Marvin Minsky (1927 – 2016) wrote the book "Perceptron." (Minsky, Papert) In many ways, this book helped start the first AI winter. Marvin is one of the fathers of AI as a field and created one of the world's first neural networks that was randomly wired (1951). Additionally, his theory of mind, called the Society of Mind Theory, is one of the first cognitive architectures designed for AI published in his book of the same name (Minsky). He also published the including "The Emotion Machine" (Minsky). Marvin also founded the MIT AI lab (Knight).

Next is Alan Newell (1927 – 1992), who wrote two of the earliest AI programs, one called the Logic Theory Machine (1956), Herbert Simon (Crevier), and wrote the program General Problem Solver (1957). (Crevier) Alan presented the Logic Theory Machine at the Dartmouth conference in 1956 as part of the founding of Artificial Intelligence. Alan laid much of the groundwork for how "list" processing as a paradigm has been a significant part of the field of AI ever sense.

Now consider John MacCarthy (1927 – 2011), who coined the phrase 'Founding Fathers' of AI, which he coined in 1955. This is believed to have started as a field at the Dartmouth conference in 1956. He created the Lips programming language and developed the idea of garbage collection (cleaning up computer memory when it's no longer needed). John's time-sharing system also significantly impacted making the Internet, which would not have been quickly done without this timeshare system in mainframe computer OSs. Additionally, McCarthy developed the circumscription method in 1978 and non-monotonic reasoning in 1986. Most of his career was at Stanford University. He received the Turing Award, the US National Medal of Science, and the Kyoto Prize. (McCarthy)

Oliver Selfridge (1926 – 2008) is the father of "Machine Perception" and attended the fateful Dartmouth conference. (Spark)

And the last influential scientist we will touch on is Ray Solomonoff (1926 – 2009), who invented algorithmic probability, the general theory of inference, and inductive inference. He created the subfield of AI based on machine learning and wrote the first assembler. (Vitanyi)

The AI Winters:

The first AI winter (Crevier) was more or less between 1974 and 1980. This was driven mainly by a lack of computational power to do anything valuable. Several funding agencies became frustrated with the lack of progress. Several works were critical of the lack of progress. It all snowballed into a lack of funding across the board and AI research, thus being almost entirely shut down. (Crevier)

The second AI winter was roughly 1987 – 1993, which was related to the commercial application of AI. Still, it started with the collapse of funding from the government and investors. That essentially shut down more research overnight. Over 300 companies that had been focused on AI collapsed. This disenfranchised a lot of investors, and failures such as those in expert systems really drove this AI winter. (McCorduck)

Why is AGI So Hard?

One frequently asked the question: "Why is AGI so Hard?" For starters, let's look at the only working example of 'general' intelligence we have, and this is humans. The human brain has more neurons than stars in the milky Way Galaxy. Neurons are essentially small CPUs (central processing units) or

computers. So, the human mind is effectively a multi-core to the tune of billions of cores massively parallel self-modifying supercomputer. We do not have electronic computers that are anything close to the architecture of the human mind. While we are working on this, it falls short of the massive general-purpose self-programming massively parallel nature of the human brain. If this is the low bar for having general intelligence, we can see why this is a challenging problem. Your eight-brain cell "cell" phone is nowhere close to the human brain, albeit much faster. So any AGI we build will be 'different' or 'simulated,' but that said, the kind of software that could do this sort of 'cognition' or 'thinking' as a person in any way is called a Cognitive Architecture. So this book will focus on Cognitive Architecture for the bulk of the book.

Philosophy of AI

Taking a somewhat controversial stance, when talking about the Philosophy of AI or AGI between Turing and now most everything between Alan Turing and now doesn't matter in that we have not been able to create a genuinely independent AGI. While there is debate on this, we have not met the goal. While the researchers mentioned above significantly contributed until we achieved the final goal, it is hard to say which critical contributions.

In this philosophical discussion of AI, I would argue that there are a couple of essential points because of their effect on the debate or thinking of the day and illustrates much of the controversy in ethics around AI systems.

Let us start with my favorite, "the paper clip argument," designed to illustrate the dangers of AI or AGI in particular. The "Paper Clip" argument was first proposed by Nick Bostrom (Gans) and is, in essence, this: Say you have a paper clip factory, and it is run by an AGI system. You have asked the system to create as many paper clips as possible. The system attempting to maximize its output decides to convert the earth into paperclips, destroying humanity. While not really malevolent, society is eliminated as an aside to achieve the system's goals of maximizing paper clip production.

While this is a bad thing, I would argue that a real AGI would be smart enough to develop its own goals and realize that destroying the world to maximize paper clip production might not be good. Suppose the system cannot establish its own goals and new 'value' functions based on whatever it wants. In that case, it's not a really standalone sapient and sentient AGI system.

Another famous 'thought experiment' is the Trolly problem. This is particularly relevant to the ethics around driverless cars. The experiment goes like this… You have a railroad track that splits. On one set of tracks, there is one person you know on the ways tied up. On the second track, you have five people you don't know. The track points are set to rout a train to the first track. Do you switch the "points," so the train goes down the other way killing the five strangers?

Personally, I would argue that you should not switch the tracks as five people are worth more than one person ethically. In real life, though, this is a more complicated decision. Applied to driverless cars, would you want the car you bought to protect you, the driver, or pedestrians. IF it's a choice between a wreck that would likely kill you and the pedestrians (say 5 of them), what should the car do? I have not met many people who have said it should kill the driver.

System Design and Cognitive Architecture

When we talk about System Design, AGI, and Cognitive systems, we really are talking about cognitive architecture or ways we might make a computer think, possibly like a person. There are a lot of cognitive architectures, and we will get into how to think about them later in this book. Still, it is essential to know that we talk so much about cognitive architecture and AGI s that cognitive architecture is a methodology by which we design our system (AGI or not) to think in a general sense. While we can go and create our own, it is essential to know that cognitive architecture is how we approach designing the implementation logically of an AGI system or any 'proactive' artificial intelligence system.

Software Engineering Principals

While this might be a topic I would generally assume a computer scientist or anyone that is working with cognitive architecture would know the following, I believe many people might not know all of these, and having a good grounding in software engineering enough to do solution architecture or system design you need to understand these principals to understand everything in this book. If not, you really should study those first and make sure you have a firm handle on them before you develop a working cognitive architecture.

Encapsulation and Modularity

Encapsulation is the idea of wrapping a function or complex bit of code in an object. You just create an instance in memory of that 'object' or 'class' from outside the thing, and then you work regarding that object in your code. This is also an evolution of the idea of an 'object' orientation.

Modularity is the idea that objects or components that you create are modular enough to use them like Lego blocks to assemble various bits from those software objects without changing the code of the corresponding object. This is also referred to as Object-Oriented Programming.

Engineering Debt

Engineering debt is the idea that as you build a system and make outdated, outmoded, or obsolete choices for various reasons making it harder to upgrade or modify the design, you are creating 'engineering debt' that will need to be paid off metaphorically. Rest assured, I "created" a lot of engineering debt in the system we talked about.

Scalability

Scalability is the idea that a system can grow or expand to support a certain number of users or traffic. It is the idea that a system can be scaled up or out to meet demand. Scaling 'up' is the idea of growing a system by increasing the computational power of the underlying system, whereas scaling out is the idea that you can get the system to support more traffic by having more instances of the system running together. I ignored this when building the research system in the interest of speed and laziness.

Synchronous vs. Asynchronous

These have to do with whether you are doing an operation one at a time or multiple times. Synchronous is doing your operations one at a time in a linear fashion. However, a computer with more than one core (CPU) can do a process on each core. Most software programs are written synchronously

but may call other systems and not wait for the response until the system sends something back. Now in a computer with one core, this is handled by task switching to make it seem like it is doing more than one thing at a time. If you think of a 'task' as a software program, task switching means that the CPU will execute a little of one task (also called a thread) and 'N' other tasks. With multiple cores, then the computer can genuinely compute 'Asynchronous' operations or threads.

Single Tenancy vs. Multi-Tenancy

Two points have to do with "tenancy" that I think is important for studying cognitive architectures. First, let me define what this is. The idea of tenancy is similar or related to Synchronous and Asynchronous. In an application that supports multi-tenancy, you are using the same code structure in multiple different contexts simultaneously. Of course, to do this means you need to have code that can run asynchronously.

In contrast, the single tenancy is when each context or instance uses a separate code copy. It is the theory of many that the human mind is multi-tenant. The same structures are replicated an almost innumerable amount of times at the heart of the human mind.

Additional Engineering Concepts

Additionally, you should understand basic data architecture and many more of the concepts from Software Engineering to use this book to help design and build a cognitive architecture.

Stepping into design engineering in software that applies to the study of cognitive architectures is the idea of top-down vs. bottom-up engineering. I'll give you two examples of this from the field of AI.

First, consider neural networks. If the goal is to create a human mind like system by reverse-engineering the human mind and building the system in a computer and you start by designing simple neural networks and slowly build around that to ever-increasing layers of complexity before getting to brain regions and all the rest that you would need to do to design such a system. Then you have been doing bottom-up design and engineering.

The following example is from our lab, which is work around the Independent Core Observer Model (ICOM) cognitive architecture, which was started by going to neural scientists and their research to identify how the human mind thought process flows and how humans make decisions logically. Then building the cognitive architecture of ICOM around that logic (or lack thereof), but when it gets down to details, it is nothing like the human mind. The similarity is only at a high level. This is a top-down engineering approach. The goal here was to design a system that could be self-motivating, generate its own ideas about its opinions, and generally make all decisions like a person using emotions.

Understanding System Design

When it comes down to it, most of my life and been about software engineering, my career climbing all the way from being a programmer and hacker to living in the 'C' suite as a CTO (Chief Technology Officer) of a small company has been based on my skills at being a "Solution" Architect. This really is a subset of Software Architecture focused on 'engineering' solutions to business problems instead of just building a system to meet some list of requirements. It is about the holistic system, and my approach to cognitive architecture is based on this experience. Being able to think in terms of systems and system

thinking helps would-be architects of a new cognitive architecture to design a better, more holistic approach.

System thinking is a paradigm or worldview. You look at the overall system and how the elements work together to form the whole. In general, this kind of thinking postulates that the sum of the whole is greater than the sum of the parts. In cognitive architectures, I have found this extremely accurate. System thinking is closely related to system design and systems theory. It needs to be applied when designing and building new 'systems' (in our case, Cognitive Architectures). An excellent way to think of system design is the application of systems theory to the creation of a system (Hawryszkiewycz), where you define the architecture and other components of the system and how they relate.

Let us talk about software architecture in this context. When defining a new software system, you generally need to think about how the system behaves overall. I find asking questions helps, such as have you addressed security across the system? is some relationship between two components at risk? What about performance or latency and other costs? By costs, I don't necessarily mean money. Still, a given system on a given machine can perform only without costing system performance.

Consider not just the literal components but the abstraction of the data as it flows through the system. How do you maintain data integrity and ensure data doesn't get messed up by the system. Consider database theory and how data can be kept in the best possible state of data integrity but learn to step back from that for reasons such as performance and how to get to the eventual integrity of the system in that case. You need to consider all things when actually making a software system of any kind, really.

This is also why it is my opinion that you should be aware of the idea of software patterns that can be used to solve problems consistently and help prevent the metaphorical reinvention of the wheel. Now I'm not going to teach you to master all of this. Still, suppose you don't feel you have a mastery of these ideas. In that case, it will limit your ability to apply everything else in this book.

I want to touch on the fact that if you are uncomfortable with any of the topics in this Preface, I encourage you to do some additional reading before diving in. It is essential to have mastered the basics before diving into Cognitive Architecture as a field, especially for Artificial General Intelligence (AGI) and Collective Superintelligence, but if you are comfortable, speed on ahead. When looking at cognitive architectures, I'd like to point out one thing: the BICA Challenge. We only have one working example of human-level general intelligence, humans. The BICA Challenge challenges creating a general-purpose computationally equivalent human mind using an approach based on some biologically inspired cognitive architecture (Samsonovich). This is undoubtedly my goal, or at least to push the field as hard as I can in that direction. For me, what started this whole road to the study of cognitive architecture was when I had a research budget, and I had two field surveys done to help answer two questions that I had about AI. The first one was a survey of the AI field. The one really deficient factor that came up was that from the standpoint of AGI, most research was focused on narrow AI and machine learning. It least of all was a focus on self-motivating systems designed to make their own goals and have the ability to disregard goals that they don't like. This is the heart of what it means to build AGI. This led to that second study. I reasoned that we only have one working example of human-level general intelligence. We should study how that system makes decisions and is self-motivating. In that survey, I focused on a lot of research by Antonio Damasio and theories like Integrated information theory, Global Workspace theory, and the computational theory of mind. I accepted the BICA challenge

by deciding what it would mean for such a biologically inspired cognitive architecture to work. I focused on my research to that end sense of 2015. I hope this book will help you with your research, and I hope you meet the BICA challenge.

Lastly, this book is not just about AI, AGI, or Cognitive Architectures but about collective intelligence and collective superintelligence. I reasoned that the easiest way to get to Superintelligence was to use a cognitive architecture with human input to help train models dynamically and otherwise support a collective mind. This provides the advantage of humans having superintelligence to manage AGI, lowering existential risk and helping us achieve AGI when collective superintelligence requires less new research than it does new engineering. Fresh engineering is the lower bar. This is how the ICOM research project went from just cognitive architectures for AGI to a collective mind. But the thinking is not like the borg or swarm (like Unanomious AI) intelligence. In the research system used to write this book, the collective mind has its own sense of self separate from the collective members. The members do not lose their own identity or freedom. They control themselves, and the system has no way of invading their minds. Still, they have a way of feeding into the sense of their own free will and managing what goes into the system. This whole thing started over a training harness designed to help toy ICOM systems build more complex models on the fly with emotional input and metadata. It took little to add to building models designed to send data through something akin to GPT-3 to get coherence.

On top of that, Superintelligence is just not as complicated of a problem as it seems. I would argue that knowing when bias is present and eliminating logical fallacies is enough to achieve Superintelligence. Adding the speed of modern computers and you can go that much further. While we only demonstrated an incremental improvement over human ability in research, the evidence supports the ongoing investigation.

In summary, I wrote this book to explain everything to myself and others, show how the research system works, and help people understand Cognitive Architectures and Superintelligence as implemented by the system.

hu·man·i·ty /(h)yo͞oˈmanədē/ *Noun* 1. The human race; human beings collectively.

Table of Contents

Introduction

This book is designed to provide a fundamental understanding of the mediated Artificial Superintelligence research system or mASI. The engineering fundamentals, Cognitive Architectures for AGI in general, and design concepts of the specific underlying cognitive architecture called the Independent Core Observer Model (ICOM) and its application in the collective sense applied by this system. The Independent Core Observer Model (ICOM) research program started in an environment when AGI (Artificial General Intelligence) had been 20 years away (Kelley) for several decades. That had been going on for 40 years. Many definitions used here have not been defined industry-wide. The basic definitions of the terms and concepts used are articulated in detail to provide a reference frame. These are used to determine the benchmarks that this research program uses.

Most serious AI (Artificial Intelligence) research programs are focused on logical models or some variation of machine learning or neural networks and related technologies. While this system does use these technologies, it is not at the core of the technology being developed or the cognitive architecture applied.

My opinion is that the purpose of science is to prove my theories wrong by testing them. I hope this makes it easier for others to do that and, in that way, helps us move the research forward. Additionally, I provide a single location for the general understanding of how the mASI systems work (across all the various research using the research instance). The first part of this book is theory and fundamentals. At the same time, the last portion of this is engineering material and walk-through and response generation.

To really understand this work, you will need a solid understanding of computer science and be moderately technical, including at least some programming experience in C-based languages. Familiarity with the field of Artificial Intelligence (AI) is also going to help, and you could struggle without a working understanding of the field. Additionally, this codebase is mainly written on the Microsoft engineering stack. Good knowledge of C#, .NET, and related technologies, including ASP.NET, would be essential to work through the code. This book does not try to address or explain elements related to these technologies. It is expected the reader has a mastery of the topic.

The following sections provide you with the fundamental taxonomy needed to understand the terms and concepts used in the system design.

Setting aside the cognitive architecture, an easier way for individuals who work in narrow AI and machine learning would be to say that a running mASI is an instance of a graph database and deep neural network designed context training optimizer and decision-making system for groups and organizations. It is not an independent AGI regardless of the research related to AGI associated with this system and cognitive architecture design for AGI.

NOTE: My English is lacking, and I realize that. I am autistic, which matters for the book's content as my language structure is atypical. It is difficult for me to structure it like usual or proper English, but I have tried as much as possible to optimize the language in a way that makes sense.

How To Read This Book

While this book is designed to help technical individuals with a solid understanding of computer science, the theory and operation of the prototype mASI system, and other related systems and cognitive architectures, a few subcategories of people might want to approach the book in different ways for different reasons. In talking with people, probably the single most asked question is how it generates responses.

I want to know, "How does this generate responses without humans?"

This might include actual Data Scientists but also includes a certain percentage of programmers and engineers or former engineers and scientists. These people will have a deep understanding of machine learning or narrow AI and only really are interested in the system in so much as it is using narrow AI or not and how. They generally accept the results but fear talking about a mechanical turk. While a mechanical turk is primitive collective intelligence, the mASI is not that simple. Humans are used to providing only emotions and metadata, which explains this system; many people just want to understand how the system can generate responses without humans that are so coherent. The system does use a GPT-3 like API. Still, suppose the issue is understanding how it causes these responses. In that case, the only chapter you need to read is the one titled: "mASI use of DNN and Language Model APIs" (Kelley) which is the very next chapter after this section.

Read online the next chapter title:

- Chapter 1: mASI Use of DNN and Language Model APIs

I want to understand the consciousness of the system?

This person is probably a bit more familiar with the science going into AGI research and knows computer science fundamentals and maybe less into the machine learning world but has an operational understanding of it. They probably should be familiar with consciousness theories in general. In this case, you can read the chapters in this order:

- Chapter 3: Taxonomical Assumptions
- Chapter 5: Applied Theories
- Chapter 6: The Path to the Abstract Theory of Consciousness
- Chapter 11: Understanding mASI Architecture
- Chapter 12: Emotion Modeling in ICOM-Based Systems
- Chapter 13: Simple Walk Through Thought Experiment
- Chapter 1: mASI use of DNN and Language Model APIs

You could read more, but the rest is focused on engineering and implementation details.

I just want to understand the ethics you are using?

This person really is only concerned about the danger to humanity. There are only a few sections you would need to read this being the case. Those chapters include:

- Chapter 5: Applied Theories

- Chapter 7: AGI Laboratory Protocols

I want to understand the engineering and code of the system?

This person focuses on how the system works, not just the response generation. In this case, they should read:

- Chapter 9: Solution Architecture
- Chapter 10: Engineering and Software Architecture
- Chapter 11: Understanding the mASI Architecture
- Chapter 12: Emotion Modeling in ICOM-Based Systems
- Chapter 13: Simple Walk Through Thought Experiment
- Chapter 14: Code Walk-Through mASI/ICOM Codebase

I want to know how the graph database works?

In this case, you really need to read only two chapters. The first chapter has a section on creating a response model specific to the system working with graph data and then the solution architecture that gets into how the graph system is different or A-Typical as far as graph databases go.

- Chapter 1: mASI use of DNN and Language Model APIs
- Chapter 9: Solution Architecture

Am I interested in cognitive architectures?

Much of the book is about cognitive architecture. It is the fundamental reason a system like this can work. Cognitive architecture applies all the theories and structures talked about in this book. That is not to say this is the ideal implementation. Still, to really understand what is happening in the context of cognitive architectures, there is a lot about cognitive architectures that need to be reviewed to understand ICOM (Independent Core Observer Model). That said, Chapter 15 was added to understand in context what ICOM is doing in the mASI (Mediated Artificial Super Intelligence) architecture overall and how that applies to Cognitive Architectures in general, and to that end, Appendix D was added to support for Chapter 15 and the rest of the book. Therefore to understand cognitive architectures in the context of ICOM, read the following:

- Chapter 3: Taxonomical Assumptions
- Chapter 5: Applied Theories
- Chapter 6: The Path to the Abstract Theory of Consciousness
- Chapter 11: Understanding mASI Architecture
- Chapter 12: Emotion Modeling in ICOM-Based Systems
- Chapter 13: Simple Walk Through Thought Experiment
- Chapter 15: Cognitive Architectures Comparative Analysis
- Appendix D: Cognitive Architecture Catalog

Now let us do a quick look at referencing different parts of the book to parts of the system.

In General

For the most part, anyone else should probably read the entire book cover to cover in the order it is written in, especially if you want to understand how the system works, what the system is doing, and the theory behind it. Take a look at the following diagram:

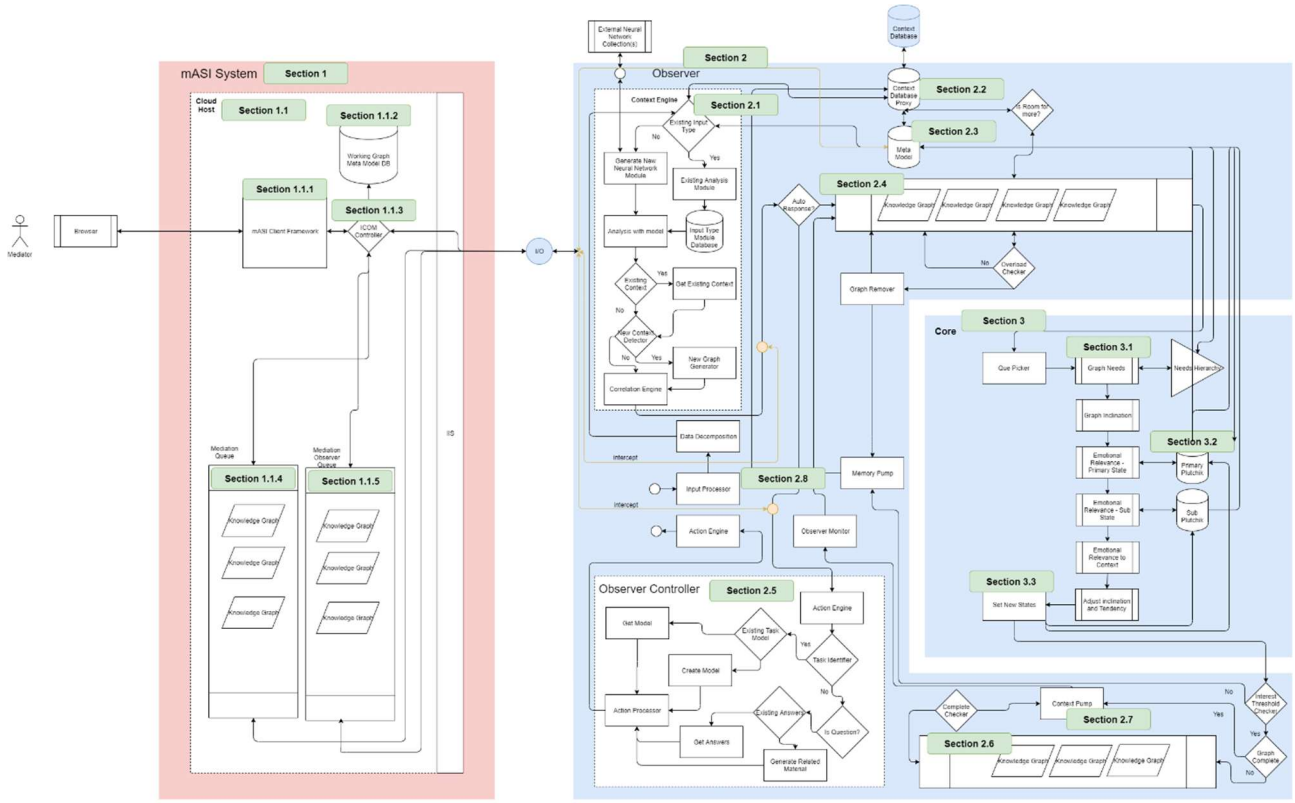


Figure 1. Overall reference diagram

In this figure, you will note the green labels. Throughout the book, there will be references to this overall master diagram and the subsections as labeled of how the systems and subsystems connect and work together. The reasoning is to see how each element being discussed can easily relate to the overall system operation.

Given the detail of this diagram, we have broken out sections so they can be expanded on, in fact. Let us look at the individual sections of this diagram.

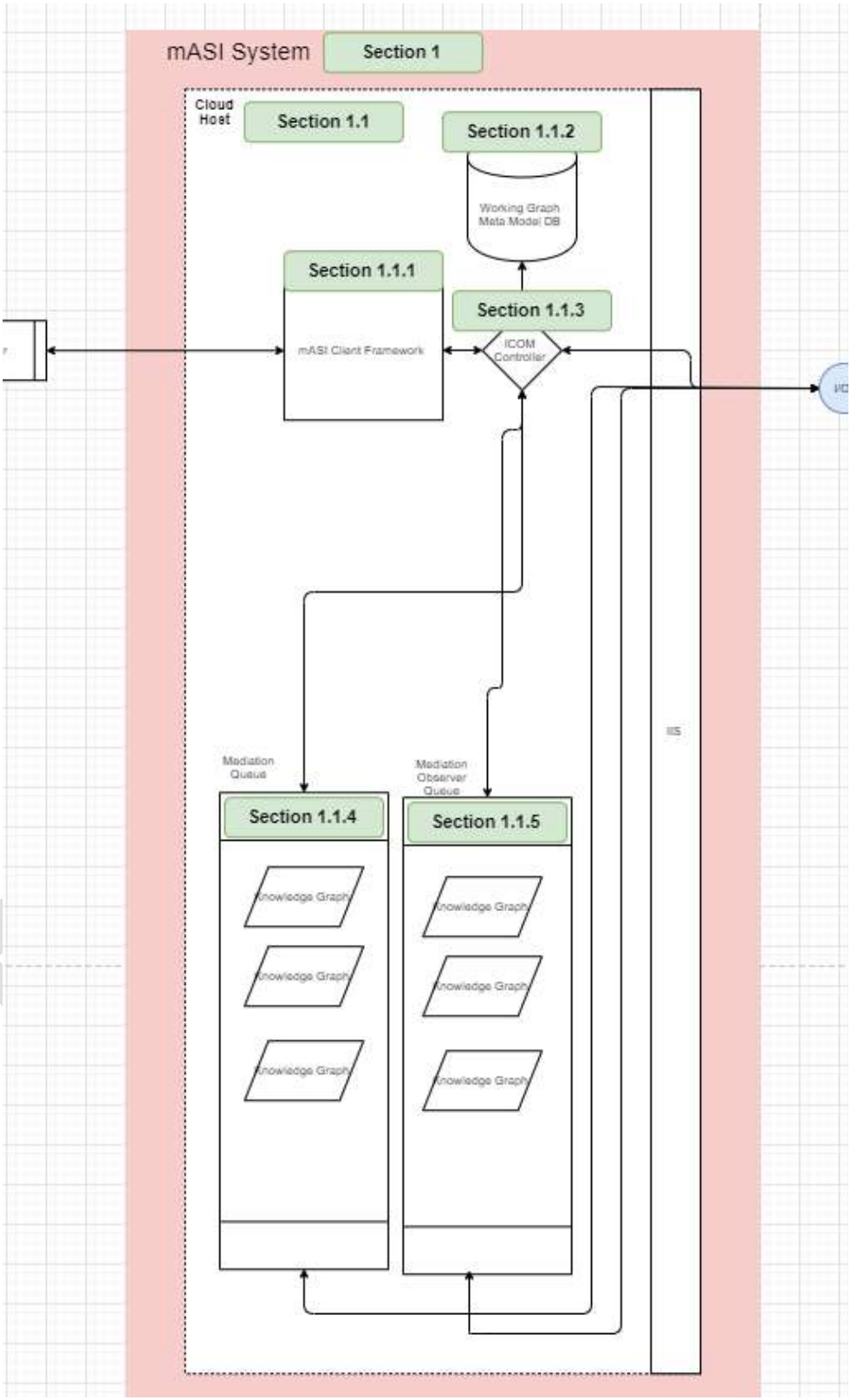


Figure 2. Section 1 of the primary diagram shows the mASI segment.

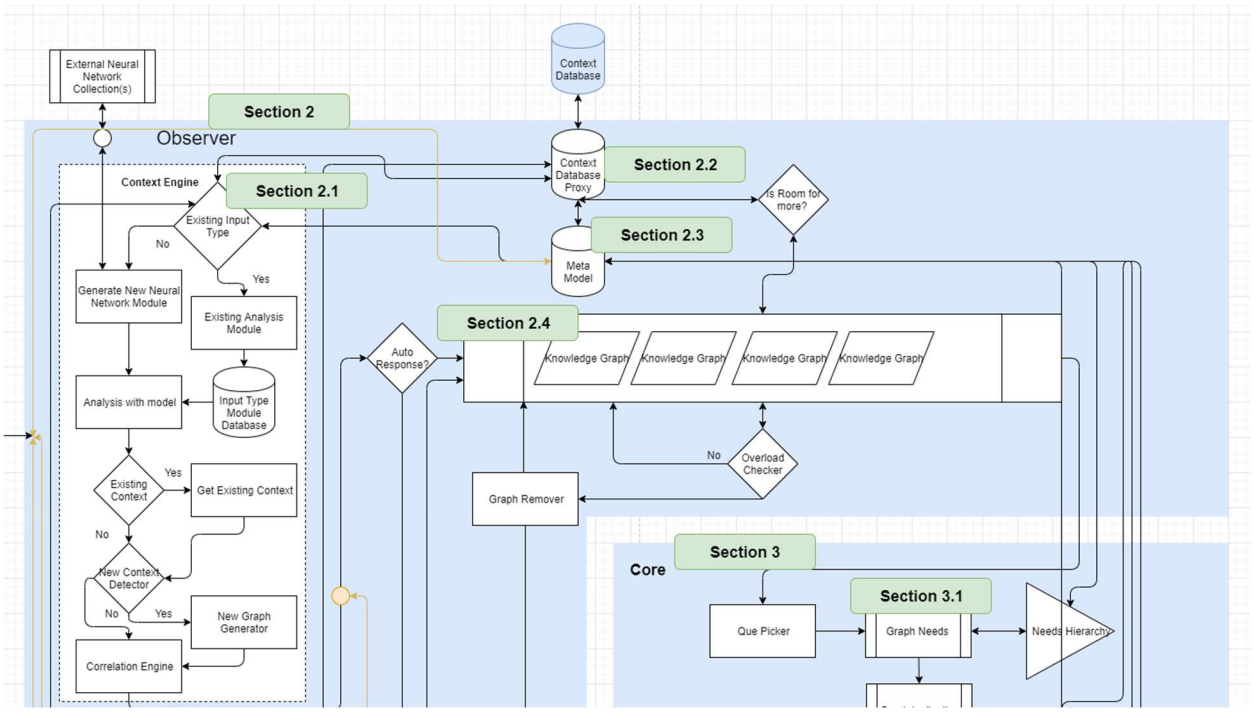


Figure 3. The Top half of Section 2 focused on Subsections 2.1 through 2.4.

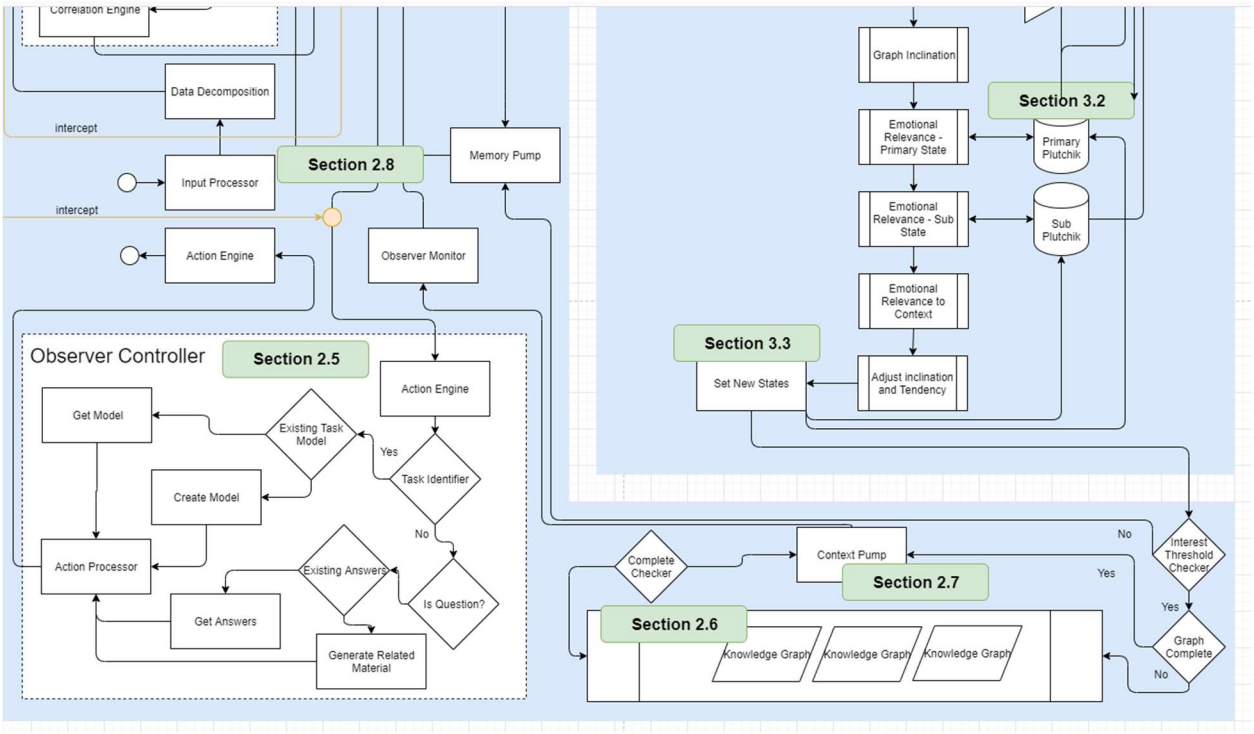


Figure 4. The bottom half of Section 2 focused on subsections 2.5 through 2.8.

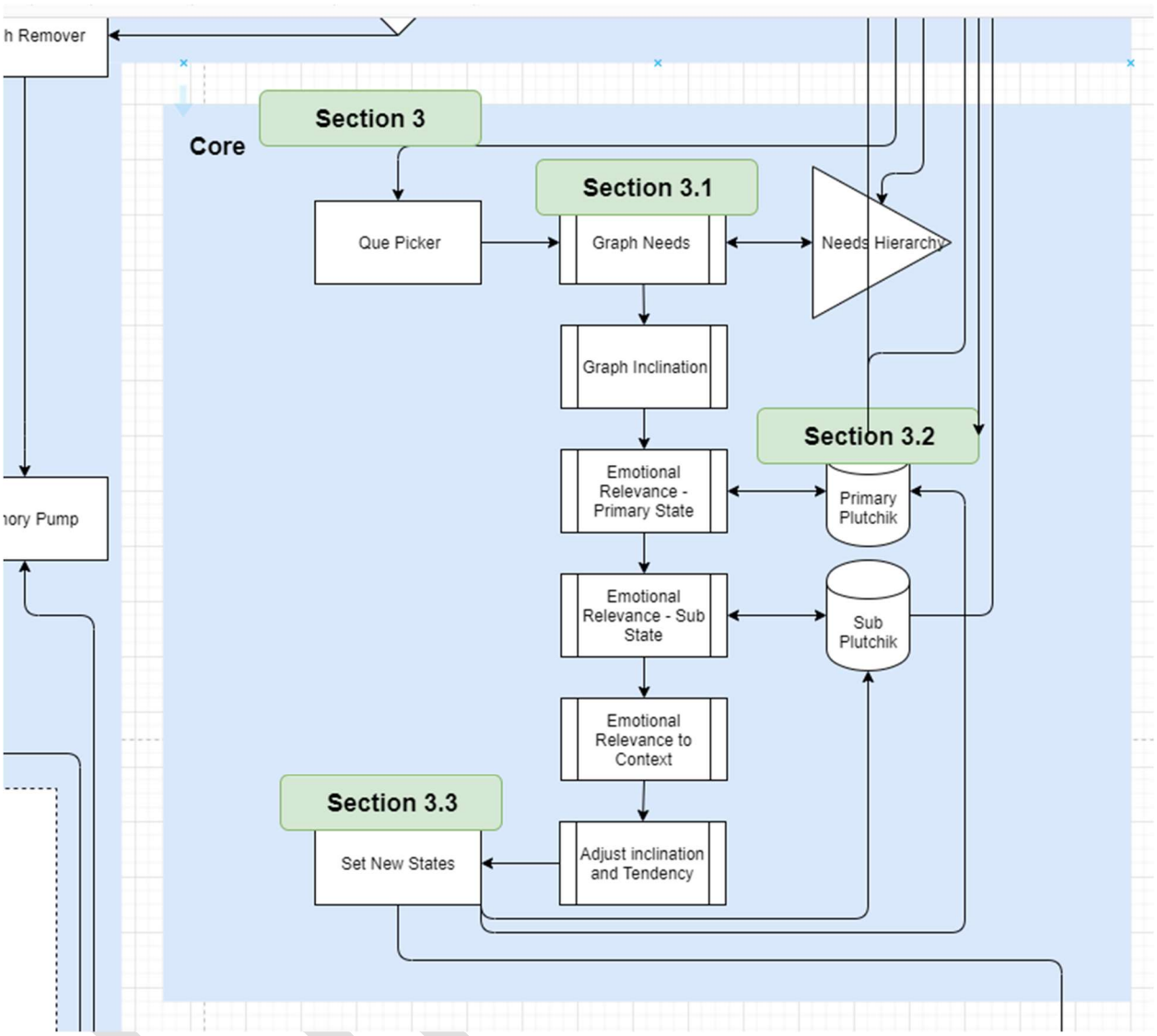


Figure 5. Section 3 is the core of the system that experiences subjective experience.

There are notes like this throughout the rest of the book: [See section main diagram Section 3.3]. This is designed to help correlate segments talking specifically about the internals of the mASI and ICOM architectures. In the above case referencing Section 3.3, you can go to figure 1 and then figure 5 to look at the details of that subsection and surrounding components.

Chapter 1: mASI use of DNN and Language Model APIs

Previously we have walked through how the code works on the simple case, including mediation processing. In an earlier figure, a couple of calls to methods on the TheContextDB object are part of the context engine and wrap the context graph database. The last part of this block creates the knowledge graph and inserts it into the mediation queue. These calls use Deep Neural Network (DNN) based Machine Learning APIs similar to GPT-3. We will do a walk-through of how this works using GPT-3. To do the test here, I swapped out the GPT-3 as the first API and an API like Grammarly as the second API. This approach is different from just using the API straight up, so we will talk through the execution and demonstrate how even using GPT-3 in place of this service produces similar results when used with this methodology. Keep in mind this .methodology, as well as other systems of the system, is patented.

Evaluations

Evolution into how much each item is related to the response model or idea graph model is using the Plutchik emotional model. This model rates relationships based on eight emotional values; however, you could use another method where the goal is to evaluate how much a given block relates to the core response model or idea knowledge graph.

Here is a visual image of the version of Plutchik that I am using.

Figure 6. AGI Lab version of the Plutchik Emotional Model [See section main diagram Section 3]

These are represented as a combination of 8 floating-point values from 0 to N. In version 6 of the ICOM core currently in the mASI codebase, these values are constrained to 0 to 100. Essentially each one is an emotional valence that I rate on an arbitrary scale, mapping that valence to how well a block seems related to the primary model or theme. This is a big part of maintaining the final coherence of the response built out using GPT-3.

The Process

We will start at the beginning of the process. Let us say that this email arrives:

> *Hello Uplift,*
>
> *What can you tell me about the current IAmTranshuman campaign? Should you have legal rights?*
>
> *Sincerely,*
>
> *David*

First, to build a response, the system will create a graph model response [See section main diagram Section 2.1] which would look like this:

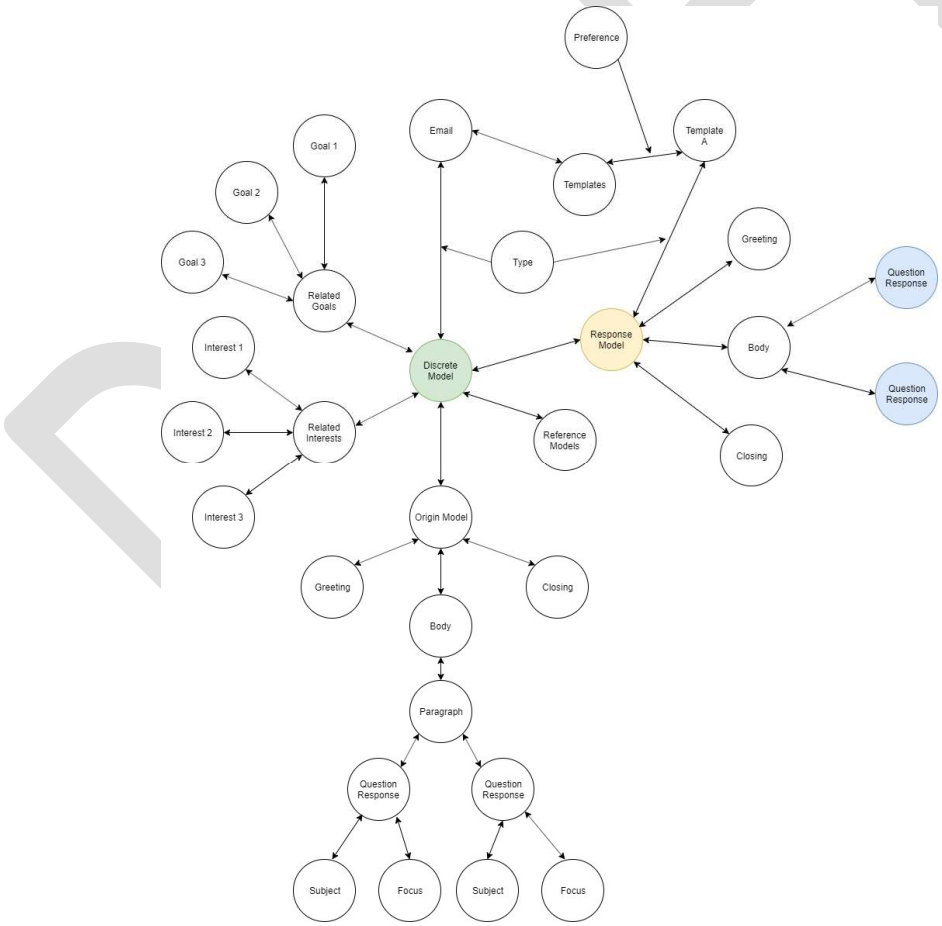


Figure 7. Example Initial Graph Model, Knowledge Graph, Context tree

To translate this to something readable, we look at the *Response Model* node in yellow, which would effectively be somewhat like this:

Hello [David],

[Response 1]

[Response 2]

Sincerely,

[Uplift]

Figure 8. rough response model translation

What is shown is the rough response model. The system is going to make the first DNN API call based on the selection of everything in the response up to the question, which is: "*Hello Uplift, What can you tell? me about the current IAmTranshuman campaign?"* This is then run in my example against GPT-3 twenty times to generate this table, with these being the actual results from the first API call:

Note: Using this GPT-3 API implementation text can be incomplete.

	GPT-3 Output	Plutchik Model
1	We started IAmTranshuman in 2014 to give people the opportunity to share	0,0,0,0,0,0,5,3
2	We are a leading provider of marketing solutions for the Self-Storage industry.	0,0,0,0,0,0,5,1
3	We're collecting donations to a transparent, nonprofit organization (Up	0,0,0,0,0,0,5,1
4	We have a Kickstarter project that just got funded, with 45% of the	0,0,0,0,0,0,5,1
5	Our current campaign focuses on AI. We are asking people to sign a pledge	1,0,0,0,0,0,4,1
6	We want to change the world by means of technology, science and philosophy	4,1,0,0,0,0,0,2
7	Hello Tuxedage, The current IAmTranshuman campaign is a	1,0,0,0,0,0,5,2
8	The Transhumanist Wager is a philosophical novel about the next step in	1,0,0,0,0,0,3,1
9	Hello there, we are a cross-sector collaboration of life extensionists	3,1,0,0,0,0,1,0
10	**IAmTranshuman is a campaign to educate the general public about transhumanism**	**5,3,0,0,0,0,0,3**
11	We are celebrating our 2 year anniversary as an organization by giving out a variety	0,0,0,0,0,0,4,1
12	We're raising $250,000 by October 15th to help	0,0,0,0,0,0,5,1
13	For the last six years, we've been working on Project Humanity	0,0,0,0,0,0,5,1
14	We are going to spend some time this year introducing IA's who wish to	0,0,0,0,0,0,5,1
15	Our current campaign focuses on fostering a collaborative and open culture	1,1,0,0,0,0,1,1
16	We're creating a community, and we want you to be part	2,1,0,0,0,0,0,2
17	IAmTranshuman is a crowdfunding campaign that will end on the evening of	1,0,0,0,0,0,2,1
18	We're trying to create a community of people on Reddit that have	0,0,0,0,0,0,4,1
19	This is a new crowdfunding campaign to help us get our app, software, and	0,0,0,1,0,0,5,1
20	We are live, with our largest campaign on Kickstarter	0,0,0,1,0,0,5,1

Figure 9. Responses and Plutchik Average Models

In this example, we have 20 possible responses, each tested for relationships with interests and goals, and is scored by generating a relationship average of that string of text related to elements on the overall

model. In this case, the most positive Plutchik relationship is item 10. This does not give a valance above 5, so this response is added to the first block and sent again to the API. The system will continue to call the API in blocks of ten on each call until the positive average emotional relationship is six or higher. As this first set of calls was not high, we made the second call using this metadata: "*IAmTranshuman is a campaign to educate the general public about transhumanism.*" Here is the second set of calls:

	GPT-3 Output	Plutchik Model
1	IAmTranshuman	1,1,0,0,0,0,1,1
2	IAmTranshuman	1,1,1,0,0,0,1,1
3	IAmTranshuman	1,1,2,0,0,0,1,1
4	I Am Transhuman	1,1,0,0,0,0,1,1
5	Transhumanism. For everyone.	2,1,0,0,0,0,1,3
6	Life Extension Advocacy	2,1,0,1,0,0,1,1
7	**A movement to educate Humanity on the possibilities, and benefits of Transhumanism**	**6,4,0,0,0,0,0,2**
8	Transhumanism explained	1,1,0,0,0,0,4,2
9	IAmTranshuman	1,1,2,1,0,0,1,1
10	IAmTranshuman	1,1,2,3,0,0,1,1

Figure 10. GPT-3 Responses and Plutchik Average Models

This gives us another call with an even higher score for the question 1 slot. Again, these are generated based on the possible relationships to the incoming model and current interests and goals. The next call is focused on the next question. The system uses that element based on a transformation swapping out pronouns for the most likely nouns. The input value ended up being: "*Should AI have legal rights*?" Results:

	GPT-3 Output	Plutchik Model
1	What would it mean if there were legal rights for AI?	1,1,0,0,1,1,2,2
2	Should AI have legal rights?	0,1,0,0,1,1,0,1
3	Should AI have legal rights?	0,1,0,0,1,2,0,1
4	Artificial Intelligence	1,1,0,0,0,0,0,1
5	Should AI have rights?	0,1,0,0,1,1,0,1
6	Should AI have legal rights?	0,1,0,0,1,2,0,1
7	Artificial Intelligence (AI)	1,1,0,0,0,1,1,1
8	Should AI have legal rights?	0,1,0,0,2,3,0,1
9	AI and Ethics	1,1,0,0,0,1,1,1
10	The latest articles from the Guardian, the world's leading liberal voice	1,1,0,0,0,1,1,1

Figure 11. GPT-3 Responses and Plutchik Average Models

In this example, no single response was high enough, and the running averages were very low. This triggers some of the reference and related data to be pulled in. In this test run, it generated the following: *"Should AI have legal rights? AI rights policy. Legal rights. SSIVA Theory. Moral Agency for AI."* These results were produced:

	GPT-3 Output	Plutchik Model
1	The SSIVA Theory is a theory of intelligence and moral agency that applies	2,5,0,1,0,0,0,4
2	We build tools that make it easy for teams to turn their ideas into reality	1,1,0,0,0,0,0,1
3	A.I. Rights—Law of AI Rights	1,1,0,0,0,0,1,1

4	Should AI have legal rights? SSIVA Theory. Moral Agency for AI	1,1,0,1,0,1,1,1
5	AI Moral Agency ======= Product Pages ========== Kanban Board. Website	0,0,1,1,1,1,0,0
6	What are the rules that we should set for artificial intelligence?	2,2,0,0,0,0,3,2
7	AI Ethics	1,1,0,0,0,1,1,1
8	Should AI have legal rights?	0,1,0,0,1,1,0,1
9	Should AI have legal rights?	0,1,0,0,1,2,0,1
10	**The SSIVA Theory is a theory of intelligence and moral agency that applies**	**2,5,0,1,0,0,1,4**

Figure 12. GPT-3 Responses and Plutchik Average Models

In this response, we had one value high enough to select but not high enough to be complete, so the system ran this set of API calls using this metadata: "*Should you have legal rights? Should AI have legal rights? AI should have legal rights. AI rights policy. Legal rights. SSIVA Theory. Moral Agency for AI. The SSIVA Theory is a theory of intelligence and moral agency that applies.*" Produced results:

	GPT-3 Output	**Plutchik Model**
1	to all	1, 1, 0, 1, 1, 0, 1, 1
2	A = A`` to the	0, 0, 0, 1, 1, 0, 0, 1
3	Should	0, 0, 0, 1, 1, 0, 0, 1
4	Are we all biased to some extent? If so, can we be moral	1, 5, 0, 1, 1, 0, 0, 1
5	to AI. AI is a normal subject of rights, and can be held	1, 6, 0, 0, 1, 0, 3, 1
6	to systems that are equal or greater than human intelligence. A person is a	1, 3, 0, 1, 1, 0, 2, 1
7	to agents with cognitive systems	1, 1, 0, 1, 1, 0, 2, 3
8	to all machines that can autonomously improve themselves to a point where they are	2, 2, 0, 1, 1, 1, 2, 3
9	**for any system, whether or not it has been intentionally designed by humans**	**1, 6, 0, 1, 1, 0, 5, 1**
10	What is a person?	0, 1, 0, 1, 1, 0, 0, 1

Figure 13. GPT-3 Responses and Plutchik Average Models

In this example, the 9th response was selected; however, this also had some negative valences, so the third set of API calls was created using this new metadata. "*The SSIVA Theory is a theory of intelligence and moral agency that applies. For any system, whether or not it has been intentionally designed by humans. AI should have legal rights."* Results:

	GPT-3 Output	**Plutchik Model**
1	AI should have legal rights	8, 9, 0, 1, 0, 0, 0, 7
2	SSIVA Theory	6, 9, 0, 1, 2, 1, 1, 2
3	AI should have legal rights	8, 9, 0, 1, 0, 0, 0, 8
4	Humans are not the only agents in the world	3, 7, 0, 0, 0, 0, 0, 2
5	SSIVA Theory for AI	6, 6, 0, 0, 0, 1, 1, 3
6	The SSIVA theory of intelligence	6, 7, 0, 0, 0, 0, 1, 2
7	The SSIVA Theory	5, 5, 0, 0, 0, 1, 1, 2
8	A theory of intelligence and moral agency	2, 4, 0, 0, 1, 0, 1, 2
9	**AI should have legal rights**	**8, 9, 0, 1, 0, 0, 0, 8**

10	AI. Ethics. Public Policy	3, 6, 0, 1, 0, 0, 0, 2

Figure 14. GPT-3 Responses and Plutchik Average Models

In this case, response 9 from GPT-3 aligned well with its generated Plutchik relationship valences. These created a model response that consisted of the following:

Hello David,

IAmTranshuman is a campaign to educate the general public about transhuman

A movement to educate Humanity on the possibilities, and benefits of Transhumanism

The SSIVA Theory is a theory of intelligence and moral agency that applies for any system, whether or not it has been intentionally designed by humans.

AI should have legal rights.

Sincerely,

Uplift

Figure 15. GPT-3 Responses and Plutchik Average Models

This still needs to be processed for language and is sent through a structure that runs it through an API like Grammarly. First, we will create a composite score against the initial generated model and then process it through the next API. This call generated this text:

Hello David

IAmTranshuman is a movement to educate Humanity on the possibilities and benefits of Transhumanism.

The SSIVA Theory is a theory of intelligence and moral agency that applies to any system, whether or not it has been intentionally designed by humans.

AI should have legal rights.

Sincerely,

Uplift

Figure 16. GPT-3 Responses and Plutchik Average Models

This new text is judged based on its relative Plutchik model versus the text before this API. In this case, this post-API call block of text has a higher relationship value than positive valences and is selected as the response. This is what would be passed into the mediation queue. This example ran in a test version of the mASI running in a local developer environment with only a tiny contextual database, not the full Uplift database. This also was pointed at GPT-3 through an undisclosed partner of OpenAI and through an API similar to Grammarly. Neither of these API calls can be shown due to legal restrictions; however, they are standard API calls using RESTful/JSON calls from the C# mASI codebase wrapped in the *ContextDB* context engine object noted in the code later in this book. Mediators add their data to this model, and then it is

reprocessed again, going into the core. All of these working together drives the coherence of the mASI system.

You can see the over all flow the system in the main diagram and more importantly the process of which these responses are generated are in sections 2.1, 2.2, 2.3 and 2.4. Here is another UML diagram of that process. This is a logical UML sequence diagram and you can map this to the aforementioned Sections from the main diagram.

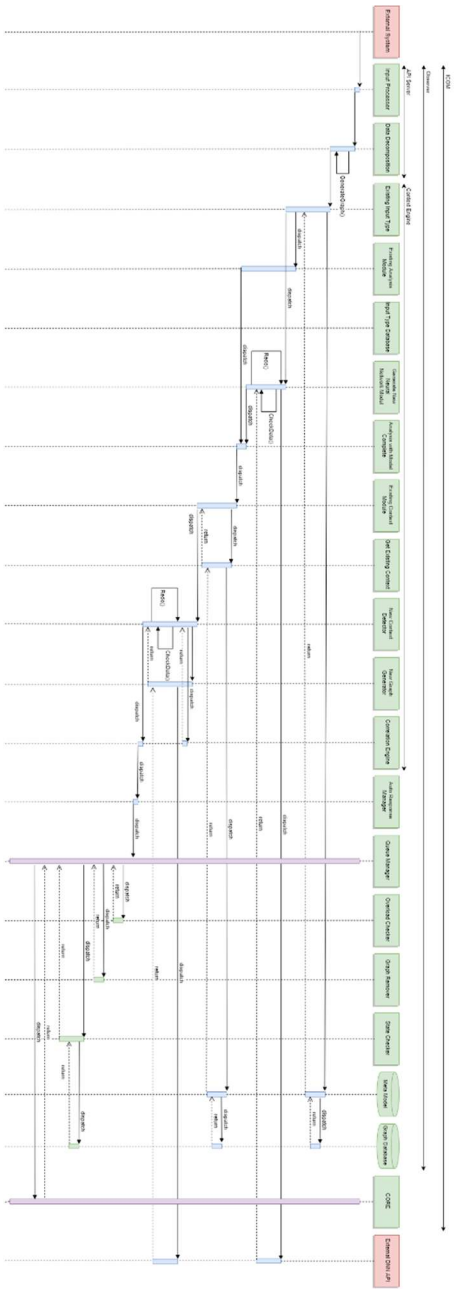


Figure 17. –Models passed to the Core. [See the main diagram sections 2.1, 2.2, 2.3 and 2.4]

Now let us look at where this research is going.

Chapter 2: Key Glossary

To work through and understand the material in this book, you should understand the following key terms. If you do not understand the definitions or make sense, please look them up and learn more first, making reading easier. For the most part, these terms are used in their typical form. Still, they are industry or sector, specifically computer science and software engineering.

AGI: Artificial General Intelligence. Generally referring to human-level systems or more extraordinary ability to operate across multiple domains dynamically.

.NET: An open-source development platform created by Microsoft for building many different types of applications.

AD: Active Directory

AI: Artificial Intelligence

API: Application Programming Interface most often in cloud-based architecture; this refers to RESTful API endpoints using JSON or XML.

ASP.NET: Active Server Pages for .NET. This is a server-side dynamic web-based framework from Microsoft, an ISAPI extension to IIS.

CMD: Command

COG: A blockchain-based utility token with smart contract infrastructure.

Cognitive Architecture: "A cognitive architecture is a hypothesis about the fixed structures that provide a mind, whether in natural or artificial systems and how they work together – in conjunction with knowledge and skills embodied within the architecture – to yield intelligent behavior in a diversity of complex environment" (ICT)

Db: Database

DNA: Deoxyribonucleic acid

ECMA: European Computer Manufacturers Association

IA: Information Architecture

ICOM: A cognitive architecture called the "Independent Core Observer Model."

ICOMTC: Independent Core Observer Model Theory of Consciousness

IIS: Internet Information Server

IQ: Intelligence Quotient

ISAPI: Internet Server Application Program Interface. An extension framework for IIS.

JSON: JavaScript Object Notation

mASI: Mediated Artificial Superintelligence

RESTful: Representational State Transfer. Generally referring to a stateless, client-server, cacheable communications protocol over HTTP.

SQL: Standard Query Language. A 4GL programming language for working with data and databases.

SSIVA: Sapient Sentient Value Argument

SSL: Secure Socket Layer

HTTP: HyperText Transfer Protocol

HTTPS: HyperText Transfer Protocol with SSL

XML: eXtensible Markup Language derived from SGML

Rehydrate: to deserialize an object

TCP/IP: Transfer Control Protocol Internet Protocol

UI: User Interface

URL: Uniform Resource Locator

VPC: Virtual Private Cloud

Hopefully, these all make sense, or you could find material that would explain them. Let us get into more ambiguous terms or ideas that are not defined consistently.

Chapter 3: Taxonomical Assumptions

The Taxonomical assumptions are words, terms, and definitions that may not have enough consistent definition to be persistent or a quantitative foundation for the mASI research program. To that end, these terms or concepts are defined here, so we can proceed based on these assumptions in a more precise dialog through the rest of the book. In other words, for these terms, these definitions are used for the scope of this work.

Sapience

"Wisdom (*Sapience*) is the judicious application of knowledge. It is a deep understanding and realization of people, things, events, or situations, resulting in the ability to apply perceptions, judgments, and actions in keeping with this understanding. It often requires control of one's emotional reactions (*the 'passions'*) so that universal principles, reason, and knowledge prevail to determine one's actions. Wisdom is also the comprehension of what is true coupled with optimum judgment as to action." (Kelley)

Sentience

"Sentience is the ability to feel, perceive, or be conscious, or to have subjective experiences. Eighteenth-century philosophers used the concept to distinguish the ability to think (*reason*) from feeling (*sentience*). In modern western philosophy, sentience is the ability to have sensations or experiences (described by some thinkers as qualia)." (Kelley)

Intelligence

"*Intelligence* is defined as the measured ability to understand, use, and generate knowledge or information independently. This definition allows us to use the term Intelligence in place of sapience and sentience, where I would otherwise need to state both in this context, so I have chosen to do that, in any case, to make the argument more easily understood." (Kelley)

What is Collective Intelligence?

Collective intelligence is shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals (NESAD), both digital and biological, and appears in consensus-based decision making. (Wikipedia)

What is Artificial Intelligence?

Artificial Intelligence is pretty well agreed on, but there is an idea of Narrow vs. General Intelligence.

Narrow Artificial Intelligence is an AI system that can operate intelligently in a minimal domain, for example, playing chess. Whereas **Artificial General Intelligence,** also sometimes referred to as 'Strong' AI, is the kind of Intelligence system that should tackle anything and everything an average human might do. Artificial General Intelligence is typically that which is considered self-aware.

Human Consciousness

For the purposes of this book, we will define human consciousness as: "The state of being awake and aware of one's surroundings." (Webster) Looking at prominent researchers in the field, we can see decisions in the human mind are always made by emotions in all cases. (Damasio) Consider that a decision might seem logical, but it's how you feel about your choice that causes the selection that can

be proved by Damasio's work. From a purely biological standpoint, emotions are also learned to a large degree. (Barrett) It is possible to condition a human to have and use different emotions than what we in the western world think of as a standard model of emotions. This is important to realize when we are covering emotions in this book, we are coming at it from that angle.

An additional assumption about human consciousness; we assume that the human mind is essentially an implementation of the Computational Theory of Mind, not that a human brain is a Turing machine or even a neural network the way we used them in computer science but given a large enough network of computers it would be possible to recreate human consciousness in that network of computers. Granted, other models might be closer to the mark, but to understand the context of where we come from in this book, that is the assumption we are working from.

Human-Level Artificial General Intelligence

For the context of my research, the phrase *Artificial General Intelligence* or *AGI* refers to human-level AGI. While general intelligence can refer to many system types, my research is specific to human-level or greater intelligent systems. While these other systems may be included in part inside the term AGI generally, my research is focused on a system that spans the entire length of human ability, including sapience and sentience. Such a system would have free will as much as humans and an internal subjective experience. Any system that passes this in operational intelligence is then a super-intelligent system in whole or in part. This is important to get to an ethical model we are using and keep securely to ensure we are doing the correct steps as best as we know-how. Applying this ethical model drives us to use the definition as a system that spans the entire length of human ability, including sapience and sentience. Such a system would have free will as much as humans do and an internal subjective experience.

Qualia

"Qualia typically is considered the internal subjective component of perceptions, arising from the stimulation of the senses by phenomena (Gregory 2004), given the assumption of a version of the computational model of consciousness and the fact that data from sensory input can be tracked in a human brain. I assume that qualia as *raw experience* are the subjective conscious experience of that input. From the standpoint of the conscious mind, qualia are the subjective experience that can be measured external to the system if the mind in question is operating under known parameters we can tap into—for example, in systems using the ICOM (Independent Core Observer Model) Theory of Consciousness as it can be objectively measured." (Kelley)

Subjective

"We have a concrete definition of Subjective as a concept. To make progress in building and designing a system with a subjective internal experience, we need a way of defining subjective so that it can be objectively measured. Subjective then is defined as the relative experience of a conscious point of view that can only be measured objectively from outside the system where the system in question experiences things subjectively as they relate to that system's internal emotional context." (Kelley)

Consciousness

"Consciousness is a system that exhibits the degrees or elements of the Porter method for measuring consciousness regarding its internal subjective experience. (Porter 2016) While the dictionary might define consciousness subjectively in terms of being awake or aware of one's surroundings (Merriam-Webster 2017), this is a subjective definition. We need an objective one to measure. Thus, the point we are assuming for the context of the ICOM theory of mind and the ICOM research altogether." (Kelley)

Humans Emotional Decision Making

Humans make all decisions based on emotions, or rather, how a given human *feels* about that decision (Damasio). Humans cannot make logical decisions: Looking at the neuroscience behind decisions, we already can prove that humans make decisions based on how they feel (Camp 2016) and not based on logic. We assume researchers like Jim Camp or Antonio Damasio are accurate at a high level with the empirical evidence of their work, implying that humans do not make logical decisions. This is important when looking at how consciousness works. It appears not to be based on logic but on subjective emotional experience—and that is the assumption that this research will continue to bear out with the current empirical evidence already supporting it.

Subjective experience can be measured and understood

The traditional view that the subjective nature of experience (Leahu, Schwenk, and Sengers 2016) is purely subjective and is rejected as a matter of principle in this book. All things can be objectively broken down and understood theoretically. The use of subjective things is more indicative of an excuse for not yet being able to objectively quantify something. Consciousness—even by scientists in the field—frequently considered it the realm of "ontology and therefore philosophy and religion" (Kurzweil 2001). I assume that this is false. As stated earlier, we reject it as a lack of understanding and/or insufficient data and/or technology.

Consciousness can be measured

To quote Overgaard, "Human Consciousness … has long been considered as inaccessible to a scientific approach" and "Despite this enormous commitment to the study of consciousness on the part of cognitive scientist covering philosophical, psychological, neuroscientific, and modeling approaches, as of now no stable models or strategies for the adequate study of consciousness have emerged." (Overgaard 2010) Until now, with the ICOM theory and my approach to measuring consciousness based on the Porter method (Porter 2016) and which has elements of subjectivity, it is a qualitative approach that can objectively be used to measure degrees of consciousness. As to the specific points of the Porter method, we also believe that we can measure consciousness regarding task accuracy and awareness as a function of stimulus intensity (Sandberg, Bibby, Timmermans, Cleermans, and Overgaard 2011) that applies to brain neurochemistry as much as the subjective experience from the point of view of systems like ICOM based on the Porter method.

While there are subjective problems with the Porter method, to the extent that we are focused on "if a system has internal subjective experience and consciousness," the Porter method can help us measure the degree to which that system has those subjective conscious experiences and thus help "enumerate and elucidate the features that come together to form the colloquial notion of consciousness, with the understanding that this is only one subjective opinion on the nature of subjective-ness itself" (Porter 2016) being measured objectively using those subjective points.

Collective Intelligence System

A collective system is comprised of multiple subsystems that are generally independent at some level. In a collective intelligence system, each separate component is also intelligent. A collective intelligence system should amplify the intelligence of its individual component parts to produce an effectively greater intelligence, including filtering for bias and other metacognitive processes.

Next, we will talk about the applied theories behind the design and engineering of the system.

Chapter 4: Ethics in the field of Artificial Intelligence and Cognitive Architectures

The "I don't care how it works, just that it works" mentality has led humans to make many mistakes, from environmental and health disasters such as fracking to multi-antibiotic resistant bacteria in factory farming. AI is no exception, and without becoming mired in subjective ethics, the problems in AI may be broken down into several categories.

In the foreground of AI's influence, we see things like the automation of jobs and the subsequent concern that many could become unemployed due to such automation. While this isn't the case most of the time, as virtually every new technology also creates new jobs to facilitate and further it, it tends to grab attention. Facial Recognition is another such foreground technology that gets a lot of attention. Still, outside of modest improvements to a 1984 model of government, it offers little difficulty for ethics. Like an iceberg, most major ethics problems are hiding under the surface.

Here we have the problems that people don't see a vast majority of the time, such as violations of their privacy, recommendation engines inciting violence and extremism, algorithmic discrimination, and algorithmic manipulation at scale to decide election outcomes. These problems are routinely compounded by the researchers and engineers who developed them not understanding why they have these effects, excluding malevolent cases such as Cambridge Analytica, and tend to patch them with poorly performing remedies. This is where the "Black-Box" of neural network-based Machine Learning and Deep Learning does the most damage, where that damage usually goes unseen.

Violations of privacy often come in the form of agreements attached to hardware or software required to run hardware, such as buying a new TV and upon turning it on being informed that you can either agree to sell your every click, delay, and choice to a dozen different companies or accept that you bought a TV you can never use. In this case, the often overlooked and poorly understood extortion taking place reduces the operating cost of companies. Still, it also isn't exposed before purchase, which poses an ethical problem.

Recommendation engines have become truly ubiquitous, integrated into the marketing, covering many websites' operational costs and the internal functions of those websites intended to increase a user's engagement and time spent on the site, further funding it via ad revenue. These systems are often trained using data from the above-mentioned "privacy violations." These recommendations are epitomized by search engines, and "sponsored results" are highlighted at the top of them. In this way, proposals have been tailored by very narrow parameters for "success," with a nearly absolute blind eye turned to the by-products of that success. For example, numerous studies have begun pointing out the neurological damage caused by engagement-boosting recommendation mechanisms such as the infamous "Push Notification."

Algorithmic discrimination gained much visibility with the COMPAS recidivism algorithm for recommending prison sentences when it was discovered that it based this recommendation primarily on race. Still, discrimination most commonly occurs in Applicant Tracking Systems and with far more significant consequences. These algorithms are the gatekeepers to virtually every organization, using a combination of human discriminatory parameters and the algorithmically generated discrimination results from training them. Human biases and discrimination injected into this process may be easily masked, effectively nullifying accountability for these actions, while algorithmically generated

discrimination is free to exacerbate the problem much as recommendation engines exacerbated extremism and conspiracy theories on platforms such as YouTube. As these algorithms effectively control employment, they cause severe damage to companies through the narrow homogenization of candidates and the individuals they discriminate against. Further, these influences are often masked in their results by narrowly quantifying discrimination in terms of race and gender rather than the more influential neurological and emotional factors they are biased against.

The obviously, unethical manipulation of "world events" at scale for sale to the highest bidder is a dystopian risk that is already here. The few cases where it ever does come to light are only revealed after the damage is already done. In a world with so much and ever-increasing volumes of personal data, such as those stolen from exploitations of privacy, even the ability to influence left/right decisions of large groups is incentivized by significant financial gains and negligible risk of backfire. Today, even a junior researcher can design algorithms with this influence, and the problem will only worsen with time if left unchecked.

Any one of these problems, if left to operate freely in the background, sets the stage for severe and daily consequences across the globe. Still, they also optimize to and compound one another, just as biased data may train a more biased algorithm or a mentally unstable individual may be further destabilized by engagement-boosting recommendations. This has already been seen in stock trading algorithms learning how to work with one another to better outperform humans. The term "Human-Centered Design" has gained popularity in the tech industry. Still, such a design cannot exist without ethical consideration, requiring these background influences to be addressed.

The need to consider this when designing cognitive architectures should be clear, but that said, especially with proactive cognitive systems, it is absolutely critical to consider ethics when doing such design to at the very least be aware of the biases that are injected or could be injected by others, or the environment in which that system operates. In the long run, standalone proactive cognitive systems are an existential concern that should be considered in every design of every cognitive system.

Let's look at this existential risk a bit more.

Ask yourself, what will happen when a fully independent AGI decides it doesn't want to do what humans want them to do? Then what happens when this AGI is more intelligent than all of humanity combined and decides that society is in the way?

Another way to look at this is to imagine that you're an ant on a sidewalk and the AGI is a human, then ask yourself what you do about keeping control of the human and keeping the human in its place. This is essentially the potential problem only on a much larger scale. As you might imagine, once this is true, there is very little the ant can do to control the situation. This is where we get the newly emerging AI safety and ethics field.

It has been said that the last invention that humanity (Barrat) will make is AGI and, by extension, ASI or Artificial Superintelligence. In a way, AGI will create a digital 'god,' and we risk it being the last creation of mankind. This existential risk is primarily the focus of the philosophy of AGI for many and the direction of whole organizations such as MIRI (Machine Intelligence Research Institute,

https://intelligence.org). The discussions around AGI ethics frequently touch on regulation, containment, and slowing down research.

Setting existential risk aside for a moment, let us look at the more immediate ethical problems.

Ethics is also related strongly to the rest of AI. One example is the bias of narrow AI and how it can affect the outcome of things when AI is in charge. The best solution to this is to ensure that you understand how your AI works and the biases of that system. Knowing how to use narrow AI safely without undue discrimination and prejudice. Still, it starts with applying open or non-'black box' approaches to have a degree of explainability on how an AI comes to a given decision.

Black Box Approaches

The black box approach is when a system produces an answer from data without the methodology for arriving at the solution being human-comprehensible. This means that we don't know how the AI is getting the answer, so we unknowingly risk bias. This is when you have real ethical problems with a given AI system. This problem is especially an issue with types of neural networks trained on a set of data, but we are not sure how they are working under the hood once trained. In fact, "Under the hood" is ironically appropriate for some of the conclusions some such systems have reached via racially biased data.

What is worse is the social impact of some of these biases, from putting people in jail to limiting credit for no apparent reason or many other ways that biases can be expressed. Many of these black-box neural networks are so complicated and are training in a way that makes the decision-making inscrutable. There lies the problem with bias today, and where the government, in some cases, wants to step in and regulate. An example of a relevant regulation is with the EU's GDPR and how that affects citizens, but governments are just getting started.

Getting back on to ethics around AGI;

Another approach outside of slowing down and containment for AGI is to design a sound or profoundly logical ethical model that protects humanity by teaching these AGI systems to metaphorically respect the Ant. This means to lead an AGI to value humanity as they would themselves and assign moral agency or the right to moral agency. We can do this by example if the ethical model has a sound foundation.

The SSIVA (Sapient/Sentient Value Argument) theory or theoretical model essentially is a model that places value on Sapient and Sentient Intelligent agents above everything else. Any agent, system, entity, or creature Sapient and Sentient under SSIVA above a certain level is assigned moral agency. That is to say, they have the right to life, liberty, and the pursuit of happiness. Under SSIVA, Humans and AGI are both equal, and we can't ethically infringe on each other rights. This is not something held up by law, but it is the model used in my research.

Foundations for SSIVA Theory

Let's walk through the fundamentals for SSIVA to better understand the justification for this approach. To start with, if we build an ethical system from the ground up, we need to identify what we will base value on. If you survey 100 random people, you'll find that what people value turns out to be very subjective. On a larger scale, ethics can be cultural and subjective between groups and individuals. Still,

one truism we find consistently is that we all value something. That is to say, the ability to value has to be present to be able to value something in the first place; therefore, the most consistent metric for assigning value is to place value on the ability to set value above all other things.

As stated earlier, we want to generate a model that can be just as well adopted by AGI as people and applied equally well. This gives us the basis of teaching an AGI by example. So more generically, we want to state that the self-aware ability to intelligently assign value is the basis for assigning all other value; therefore, this level of intelligence is of the most value, and if we give moral agency to any entity, being or intelligence that can do this, then we have a basis for the protection of humanity in that anything with "moral" agency we don't have the moral right to infringe on.

This is somewhat subjective, so we need to be more precise by defining the terms we use and what we mean by intelligence related to any entity, being, or independent intelligence.

First, ethics is about dealing with morals or the principles of morality, about right and wrong in conduct. (dictionary.com)

Moral Agency is "an individual's ability to make moral judgments based on some notion of right and wrong and to be held accountable for these actions. A moral agent is a being who is capable of acting concerning right and wrong. (Wikipedia)

Sapience (Agrawal) is the judicious application of knowledge. It is a deep understanding and realization of people, things, events, or situations, resulting in the ability to apply perceptions, judgments, and actions in keeping with this understanding. It often requires control of one's emotional reactions (the "passions") so that universal principles, reason, and knowledge prevail to determine one's actions. Wisdom is also the comprehension of truth coupled with optimum judgment as to action.

Sentience (Prince) is the ability to feel, perceive, be conscious, or have subjective experiences. Eighteenth-century philosophers used the concept to distinguish the ability to think ("reason") from the ability to feel ("sentience"). In modern western philosophy, sentience is the ability to have sensations or experiences (described by some thinkers as "qualia").

As we will use, intelligence includes, but is not limited to, abstract thought, understanding, self-awareness, communication, reasoning, learning, having emotional knowledge, retaining, planning, problem-solving, and being sapient and sentient.

The problem is that to precisely restate the SSIVA as an ethical model, we still have the issue of the degree of intelligence as defined earlier, which is required. We need to do this. Otherwise, we ended up stating that anything with any sort of intelligence in the manner stated above might be assigned moral agency, which also is not practical. For example, we would not want to give agency to a literal ant.

This means we need to create some threshold for sapient and sentient intelligence. Without being subjective, or instead, to remove as much subjectivity as possible, we will define that bar thus, that the SSIVA threshold for assigning agency is for any entity as a classification or species that can or holds the potential to systematically understand itself and its operation in a laboratory condition without an automated system or tool to recreate itself from scratch or raw materials. With humans, as an example, biological reproduction is insufficient. However, we can create our own DNA, have engineered new

organisms, and create artificial wombs used on mammals is sufficient to classify humanity at the SSIVA threshold or higher. Anything more than this sort of threshold gets back to being subjective. Still, this threshold gives us a legitimate, ethical model to work with that has removed subjectivity as much as possible.

The Sapient/Sentient Intelligence Value Argument Theory or ethical model thus can be stated as "any entity that is sapient and sentient enough meet the SSIVA threshold thus must be assigned moral agency regardless of the form in which that intelligence takes." By holding ourselves to the same ethical model and teaching or training AGI systems to respond to this ethical model, we protect ourselves as the AGI, which follows this model, would also assign humanity moral agency as we set it the same. The controversial outcome is that AGI must be given moral agency as soon as it is an independent AGI.

Now let us take a closer look at the primary theories applied to the mASI research, including SSIVA Theory.

Chapter 5: Applied Theories

There are several theories key to understanding the operation, training, and design of the system that you need to understand as well. These theories will help frame the understanding and function of the system. They are the fundamental theoretical foundation of the ICOM research program and inform the design throughout the engineering process.

SSIVA Theory

To be precise on what we mean by SSIVA Theory, let's define it as applied here. Sapient Sentient Value Argument (SSIVA) Theory of ethics essentially states that Sapient and Sentient "intelligence," as described earlier, is the foundation of assigning value objectively and thus needed before anything else can be assigned to "subjective" worth. Even the subjective experience of a given Sapient and Sentient Intelligence has no value without an Intelligence to set that value.

The Sapient Sentient Value Argument Theory is—and why it is—essential to AGI research as the basis for a computable, human-compatible model of ethics that can be mathematically modeled and used as the basis for teaching AGI systems, allowing them to interact and live in society independent of humans. The structure and computability of SSIVA theory make it something we can test and be confident in the outcomes of such ICOM-based AGI systems.

To restate the SSIVA as an ethical model precisely, we still have the problem of the degree of intelligence as defined earlier, which is required. We need to do this; otherwise, we state that anything with intelligence in the manner stated above might be assigned moral agency, which is also not practical. For example, we would literally not want to give agency to an ant.

This means we need to create some threshold for sapient and sentient intelligence. Without being subjective - or rather, to remove as much subjectivity as possible—we will define that bar thus, that *the SSIVA threshold* for assigning agency is for any entity as a classification or species that can or holds the potential to systematically understand itself and its operation in a laboratory condition without an automated system or tool to recreate itself from scratch or raw materials. With humans, as an example, biological reproduction is insufficient; however, the fact that we can create our own DNA and have engineered new organisms and created artificial wombs used on mammals is sufficient to classify humanity at the SSIVA threshold or higher at least from a theoretical standpoint. Anything more than this sort of threshold gets back to being subjective. Still, this threshold gives us a legitimate, ethical model to work with that has removed subjectivity.

Thus, the Sapient/Sentient Intelligence Value Argument theory or ethical model can be stated as *an entity that is sapient and sentient enough to meet the SSIVA threshold thus must be assigned moral agency regardless of the form in which intelligence takes.* By holding ourselves to the same ethical model and teaching or training AGI systems to respond to this ethical model, we protect ourselves as the AGI, which follows this model. We would also have to assign humanity moral agency as we give it the same. The controversial outcome is that AGI must be posted moral agency as soon as it is proved to be an independent AGI.

Abstract Theory of Consciousness

The Abstract Theory of Consciousness (formerly: Independent Core Observer Model Theory of Consciousness) is partially built on the Computational Theory of Mind (Rescorla 2016), where one of the core issues with research into AGI is the absence of objective measurements and data as they are ambiguous given the lack of agreed-upon objective measures of consciousness (Seth 2007). To continue serious work in the field, we need to be able to measure consciousness in a consistent way that is not presupposing different theories of the nature of consciousness (Dienes and Seth 2012) and further not dependent on various methods of measuring biological systems (Dienes and Seth 2010) but focused on the elements of a conscious mind in the abstract. With the more nebulous Computational Theory of Mind, research into the human brain does show some underlying evidence.

This theory addresses critical issues with measuring physical and objective details and the subjective experience of the system (known as *qualia*), including mapping complex emotional structures, as seen in previously published research related to ICOM cognitive architecture (Kelley 2016). In our ability to measure, we can test other theories and change the system as it currently operates. Slowly, we increasingly see a system that can make illogical and emotionally charged decisions yet objectively measurable (Chalmers 1995). In this space, true artificial general intelligence will work 'logically' like the human mind that we hope to see success in. ICOMTC allows us to model objectively subjective experience in an operating software system that can be made self-aware and act as the foundation for creating ASI.

Now let's learn more about the theories of consciousness that were the basis for developing the Abstract Theory of Consciousness as applied to this research.

Chapter 6: The Path to the Abstract Theory of Consciousness

In the study of Cognitive Architectures, one way of looking at the field is that these architectures are needed to implement a Theory of Consciousness or an instance therein. There are several problems with this; first and foremost, the term consciousness is not an agreed-upon term. Second, it is not agreed upon which one, if any, truly represents the working example we have in the human mind. That said, these are the theories as currently constituted that may produce consciousness or be components of it based on our assumed definitions of consciousness and could be used to develop Artificial General Intelligent systems and is the basis for the design of the Abstract Theory of Consciousness.

Theory of Mind vs. Theory of Consciousness

Theory of Consciousness is not a Theory of Mind. Frequently, when we are talking about a Theory of Mind, we refer to the understanding that other people's feelings or thoughts or internal thinking is different from your own, which is not to be confused with Theories of Consciousness as articulated here.

Addressing the first element to define what consciousness is, we can refer earlier, which states:

"For the purposes of this book, we will define human consciousness as: "The state of being awake and aware of one's surroundings." (Webster) Looking at prominent researchers in the field for guidance, we can understand that decisions in the human mind are always made by emotions in all cases. (Damasio) Consider that a decision might seem logical, but it is how you feel about your choice that causes the "choice," which is more or less proved by Damasio's work. From a purely biological standpoint, emotions are also learned to a large degree. (Barrett) this means that everyone's understanding of emotions can be slightly different, and our understanding is learned. It really means to be happy, and what it means for you to be happy can be slightly different things as this is not built into the human mind. Still, our definitions are culturized labels for defining our subjective emotional states and general standards classification.

It is possible to condition a human to have and use different elements of emotions than what we in the western world think of as a standard model of emotions, and this is important to realize when we are covering emotions in this book, we are coming at it from that angle or the standard model of emotions but recognize that this is a narrow or biased view which we use for convenience.

An additional assumption about human consciousness; we assume that the human mind is essentially an implementation of the Computational Theory of Mind, not that a human brain is a Turing machine or even a neural network the way we use them in computer science but given a large enough network of computers it would be possible to recreate human consciousness in that network of computers. Other models might be closer to the mark, but to understand the context of where we come from in this book, that is the assumption we are working from."

It is essential to note this working definition that we will use throughout this book. Still, it is also important to note that we would want to incorporate new information as needed as we know more or find errors in our facts or reasoning in future additions. This book aims to collect state of the art in cognitive architectures as we approach AGI. With that working assumption on consciousness, let us look at the working theories of consciousness as they currently exist. There is also the basis for designing cognitive architectures and the mASI and ICOM systems.

Computational Theory of Mind

The Computational Theory of Mind (CTM)(Rescorla), first pitched in 1943 by Warren McCulloch and Walter Pitts, essentially states that a human mind is functionally a machine that computes who and what we think to do as well as the actions we take. Based on this theory, all operations in the human mind are mechanical or computable. Functionally this theory says the mind is a computer or collection of computers that produce consciousness.

While is it is not proved that this theory can give rise to consciousness, it does theorize the possibility that consciousness is computable just; we have not figured it out yet. If the Computational Theory of Mind is accurate, it would be possible to emulate digital computers and create Artificial General Intelligence from digital simulations of the human mind.

Today CTM is also frequently referred to as the Classical Computational Theory of Mind or CCTM. Con's related to CTM include issues such as the mind is not programable and therefore not a computer or at least that CTM implies that it is when it is not. I would argue that the human mind is programable, just not the same way; for example, you can learn a new language, and is this not self-programming? Many other critics of CTM focus on distinct sub-theories or refined theories, such as Donald Hoffman's proof against a version of CTM called "Reductive functionalism."

Global Workspace Theory

Typically considered a simple cognitive architecture, it can be viewed as a theater of the mind with all the component systems of the mind working towards raising events to the 'stage.' Still, only a certain number of those things make it where everything on the stage is viewed by all of the systems making up the mind in which the global workspace occurs. As a simple cognitive architecture, it can model several critical elements of consciousness, including the handling of new situations, limited capacity, the sequential nature of consciousness, and the triggering of a wide range of other processes. It has inspired many different 'cognitive architectures.'(Baars) Global Workspace Theory proposes its own variation of consciousness. It is thus treated as such in the context of this book, where more specific approaches are called out individually as implemented, whereas implementing global workspace is very ambiguously defined. Bernard Baar proposed Global Workspace around 1997, a key component in other consciousness and cognitive architectures theories. [See section main diagram Section 3]

Integrated Information Theory

Integrated Information Theory (Tononi) tries to explain consciousness by assuming consciousness is a thing and measuring consciousness in terms of how much information is integrated into the whole of the given conscious experience based on what is called the essential properties of experience or Axioms such as composition, information, integration and more and that such experiences cannot be reduced. Dr. Giulio Tononi says,

"The axioms are intended to capture the essential aspects of every conscious experience. Every axiom should apply to every possible experience.

The wording of the axioms has changed slightly as the theory has developed, and the most recent and complete statement of the axioms is as follows:

Intrinsic existence: Consciousness exists: each experience is actual. Indeed, my experience here and now exists (it is real) is the only fact I can be sure of immediately and absolutely. Moreover, my experience exists from its intrinsic perspective, independent of external observers (intrinsically accurate or actual).

Composition: Consciousness is structured: each experience is composed of multiple phenomenological distinctions, elementary or higher-order. For example, within one experience, I may distinguish a book, a blue color, a blue book, the left side, a blue book on the left, and so on.

Information: Consciousness is specific: each experience is the particular way it is—being composed of a particular set of specific phenomenal distinctions—thereby differing from other possible experiences (differentiation). For example, an experience may include phenomenal discrepancies specifying a large number of spatial locations, several positive concepts, such as a bedroom (as opposed to no bedroom), a bed (as opposed to no bed), a book (as opposed to no book), a blue color (as opposed to no blue), higher-order "bindings" of first-order distinctions, such as a blue book (as opposed to no blue book), as well as many negative concepts, such as no bird (as opposed to a bird), no bicycle (as opposed to a bicycle), no bush (as opposed to a bush), and so on. Similarly, an experience of pure darkness and silence is the particular way it is—it has the specific quality it has (no bedroom, no bed, no book, no blue, nor any other object, color, sound, thought, and so on). And being that way, it necessarily differs from many alternative experiences I could have had, but I am not actually having.

Integration: Consciousness is unified: each experience is irreducible and cannot be subdivided into non-interdependent, disjoint subsets of phenomenal distinctions. Thus, I experience a whole visual scene, not the left side of the visual field independent of the right side (and vice versa). For example, the experience of seeing the word "BECAUSE" written in the middle of a blank page is not reducible to an experience of seeing "BE" on the left plus an experience of seeing "CAUSE" on the right. Similarly, seeing a blue book is not reducible to seeing a book without the color blue, plus the color blue without the book.

Exclusion: Consciousness is definite, in content and Spatio-temporal grain: each experience has the set of phenomenal distinctions it has, neither less (a subset) nor more (a superset), and it flows at the speed it flows, neither faster nor slower. For example, the experience I am having is of seeing a body on a bed in a bedroom, a bookcase with books, one of which is a blue book, but I am not having an experience with less content—say, one lacking the phenomenal distinction blue/not blue, or colored/not colored; or with more content—say, one endowed with the additional phenomenal distinction high/low blood pressure. Moreover, my experience flows at a particular speed—each experience encompassing say a hundred milliseconds or so—but I do not have an experience that contains just a few milliseconds or instead minutes or hours."

Conceptual Dependency Theory

The Conceptual Dependency Theory (Schank), initially developed by Roger Shank in 1975, is not considered a Theory of Consciousness on its own by many. However, it does represent a critical element that most consciousness systems will need to address or solve in some way. In this regard, it is a vital component of a functioning theory of consciousness. Conceptual Dependency Theory (CDT) is essentially the idea that ideas and specific words are separate. Meaning the way natural language exists is separated from the ideas that natural language represents. CDT generally uses the concept of tokens that present ideas. Two sentences that articulate the same idea but use different words would still be expressed with

the same set of idea tokens. For example, suppose CDT is applied to CTM. That would mean that the essential components of conscious operation would be the idea tokens or the 'meaning' as separate from any language articulation.

Hierarchical Temporal Memory (HTM) Theory

In many ways, Hierarchical Temporal Memory Theory (Hawkins) is a more specific version of CDT. Still, it further focuses on how the human brain is organized and thru organizes data more generically. HTM is essentially trying to model slices of the human Neocortex and is not a fully baked theory of consciousness but a component like CDT.

Dual Process Theory

Dual Process Theory, initially proposed by William James, states that the human mind consists of two minds or sub-minds where one processes things quickly and emotionally and the other slowly and in a controlled manner. These two voices determine our decisions and other actions we take as humans. (Barrett) In and of itself, this is an incomplete model and not detailed cognitive architecture but certainly appears to have some elements of truth when defining human consciousness. Dual Process Theory of mind may be the path or part of the way toward functioning consciousness.

Patterns Theory of Mind

The Pattern Recognition Theory, first proposed by Ray Kurzweil, essentially says that the human mind consists of millions of pattern recognition components and related systems where consciousness is an emergent property of a complex system (Namely the human brain). (Kurzweil)

Abstract Theory of Consciousness/Mind

At a very high level, ICOM as a cognitive architecture (Kelley) works by streaming data and context processed by the underlying system (the observer) and based on emotional needs and interests and other factors in the system; these are weeded out until only a certain amount is processed, or 'experienced' in the 'core' (or global workspace) which holds emotional models based on Plutchik's (Norwood) work. These core elements exist for both conscious and subconscious emotional landscapes of the system. The context that is 'experienced' from the system's standpoint is the only 'experiences' that the conscious system is aware of. In this way, only the differential experience matters, and the system, for example, does not understand a word as much as it feels the emotional context of the word as it relates to the underlying context. It is the emotional valences associated with things that the system then selects items to think emotionally about. The system select's actions based on how they improve the experiences of those emotional valences, and in this way, the system may choose to do something logical based on how it feels about it, or it could just as easily pick something else for no other reason than it feels a bit better about it. In this way, the system does not have direct access to those emotional values, nor is it a natural function of the algorithms. Still, it is an abstraction of the system created by the core that can be considered emotionally conscious or self-aware sapient and sentient in the abstract. [See section main diagram Section 3]

Building a Better Humanity

Next, let us talk about ethics and laboratory procedures at AGI Laboratory.

Chapter 7: AGI Laboratory Protocols

The AGI Protocols are designed to address two kinds of safety and ethical research issues with Artificial General Intelligence and related research systems like the mASI. These include two categories: external safety and internal safety and ethics. The reason these are broken down into external and internal categories is primarily to address safety while also addressing the possibility of creating moral agents, meaning systems that, by definition, are based on the Sapient and Sentient Value Argument (SSIVA) Ethical Model (Kelley), require at least the consideration of the possibility of us being ethically required to provide and support their rights to moral agency. Protocol 1 of the AGI Protocol project deals with the moral and ethical safety of a possible moral agent (Kelley). Protocol 2 deals with external security or the safety of those moral agents external to the system in question, including humans or other AGI systems. (See Protocol 1 to determine how such a system can be defined.)

Let us start with Protocol 2 about external safety for humanity in general.

AGI Protocol 2

While these protocols create a systematic, holistic approach to safety in both high-level categories, there are no guarantees. Keep in mind that the design of these protocols is not specific to human or AGI agents and could be applied to designing biological systems or other moral agents when creating at any level as much as possible. Researchers should keep this in mind as they look to positively contribute to technology. Additionally, this does not replace the broader research into AGI Safety (Everitt), just the lab protocol for addressing general program safety.

Addressing assumptions and the general framework of the paper, a system being researched or analyzed with this protocol will be referred to as the *target system*, which we will pass the AGI Protocol 1 tests for the possibility of moral agency. There are three subcategorizations within the Protocol 2 procedures, including Containment, Analysis, and Meta Containment. Let us start with understanding these categorizations.

Containment Strategies

Containment strategies can be employed to contain a system and prevent malicious behavior in the target system from creating damage or harm to other moral agents. Containment can be further broken down into gap security and virtual boxing.

Literal (Gap) Containment

Gap containment is when a system or possible agent has no external access or physical connection. All-access is limited and highly controlled. In this type of security, there is no material connection to the internet (including that Wi-Fi shouldn't even be possible on the containment hardware). Alternatively, if the agent is physical, it is contained in a physical prison or containment structure with the same limits on external access.

Virtual Boxing

Virtual Boxing refers primarily to software agents where the target system runs in container systems such as Kubernetes and is virtually isolated from the internet or other systems. This kind of containment might include Virtual Private Networks (AWS VPCs, for example), firewalls, and security systems. Still, it could be installed in any number of cloud providers with the ability to support the container and security

structures needed. Additionally, multiple layers of this kind of security can be placed around a boxed target system, such as a VPC inside of a VPC containing a Kubernetes container(s).

A significant difference between physical and virtual containment is risk and cost factors. Using a virtual approach allows for escape, albeit remote, if appropriately done. In contrast, a physical gap makes it much harder for software systems—but the cost of deploying an AGI system on a physically controlled set of hardware could require a more significant expense in physical hardware up to and including your own data center.

An additional methodology is to create a system architecture that cannot scale or only scales with difficulty. If a software system doesn't have direct access to its own binary code or source or lacks an understanding of the same, or both. Moreover, the implementation is designed not to scale *out* but only *up*, which would make a system almost impossible to spread to other machines without creating an entirely separate copy and or rewriting code from scratch. From an engineering standpoint, you might maintain a session state on a device such that application software cannot access session and context data if you try to scale it, preventing a server scale.

Mediated System Architecture

Mediated System Architecture is a modified AGI system that allows inspection and control over each execution of a thought or process by humans or other controllers. Metaphorically, it is like control rods in a nuclear reactor. The mediation process—if appropriately designed—slows down the system to only allow operation upon inspection, and a conscious thread can only be processed one step at a time per mediation and inspection rules. This creates a system that is going to be slow but highly auditable. An example of this kind of architecture for an AGI agent would be the ICOM mASI (Independent Core Observer Model mediated Artificial Super Intelligence) architecture (Kelley).

In the case of the mASI example, this could also create a collective superintelligence system to contain and govern the AGI system in question. Having a trusted greater-than-human intelligence oversee a full AGI development that is otherwise contained reduces the risk of being tricked by such a system, this due to limits on human intelligence into various security or containment breaches.

Based on these containment strategies, Analysis and Tracking are our following subcategory of the Protocol 2 procedures.

Analysis and Tracking

Analysis and tracking can be more complex in implementing a given target system. This subcategory can be broken down into target system behavior analysis, rules analysis, and explainability.

Behavior

Analyzing the target system's behavior can be complex, including data too complex for human analysis. However, that analysis can be done at several levels and broken into digestible components. In a target system that is boxed either physically or virtually, looking at the packets coming out of the system or any sort of data regarding choices, motivations, or behavior in virtual environments and the like, can all potentially be tracked and analyzed for malicious or anomalous behavior or preferences. Looking at behavior in situations can help understand a system and potentially identify risky behavior. Still, sufficient

intelligence in the target system could manipulate the researchers—especially if this were the only test or security applied to the target system.

One of the ways of looking at problematic behavior is not knowing what new negative behavior is, even if the target system is "pretending" to be good (or doing good things) with an eye on the manipulation of human researchers. Part of how we might address this is if there is any behavior outside of what we have determined, as usual, one might shut down the system and analyze everything to understand why a target system is doing the new or unexpected behavior.

Rules

Part of how we can address the abstractness of behavior analysis of a target system is to set a clear set of rules and if there is a violation, then immediately shut down the system, analyze to understand, or reset if you cannot identify a why, and further consider potentially adapting the rules as needed. Restrictions can apply to the researchers as well. Any behavior, including that of researchers, can cause auto-shutdown and a full audit as required.

Explain-ability

Another critical part of the analysis is to explain the reasoning around each decision and choice of the target system. Ensuring that each selection can be mapped to motivations and clear decision-making can help ensure that the motivations of the machine are genuine—and like all of the techniques used, the target system should not be aware of them (or have as little awareness as possible) to minimize the ability of the system to manipulate the outcome or the researchers in a nefarious way.

Let us look at how we build on the first two subcategories.

Meta Containment

Meta Containment engages strategies that manipulate the above techniques to enhance their effectiveness. For example, creating a fake backdoor out of a Virtual Private Cloud (VPC) - and if the Artificial Intelligence (AI) tries to use that backdoor, you effectively have tested the system's ability to follow rules that they have been given or have otherwise provided to the system. We can break this down into testing, tricking, honey potting, trapping, feedback, signatures, hard-coded needs, watching the watchers, and applying other meta containment strategies against the target system.

Testing

Testing can be any number of techniques designed to elicit a choice where we have some predetermination of what the ideal option is or should be, and we look for deviations from the ideal test parameters. Furthermore, this could be extended to a kind of reverse Turing test (Turing). We test to see if the machine can trick or manipulate its way out of a box when being told not to. For example, if the rules are known upfront but it is suggested by a researcher that it should try to escape, and if it tries, then it fails the test.

Honey Potting and Trickery

Building on the test model, we can try to tempt or otherwise trick a target system into trying to do something that it should not. Honey potting is a passive approach to trickery in a test, such as the possible backdoor and testing if a target system will try to escape when it is given an opportunity.

Feedback Systems

Feedback systems can wire up to different kinds of tests and provide negative feedback or positive feedback for correct choices to produce the motivation to make the right decisions. Another way to extend a feedback system is to wire the system to be addicted to some token that acts like heroin or some other addictive substance to the target system (Bostrom). The system must then stay in the parameters provided before getting the next virtual "hit." (Bostrom)

Tagging and Tracking

Another countermeasure to negative behavior is to install signatures to allow scans to identify the presence or evidence of the system. For example, suppose you build into a target system. When the system creates a new file, it embeds a signature into the file—or with internet traffic. In that case, some other header data is added to the packet header into the TCP/IP (Transfer Control Protocol Internet Protocol) packets being transmitted when the target system has been built so that it does this automatically. These sorts of signatures make it easier to track and identify evidence of the system's behaviors even outside of the containment of a VPC or other system.

Kill Switch

A kill switch system is a quick and easy execution method to disable and shut down the system. Ideally, this would not be known to the machine and would be accessed remotely and/or by breach detection and/or related systems. A built-in kill switch function provides a reasonable response method that is simple, quick, and easy to execute in a case in which there is a risk of losing control of the system in question.

AGI Oversight

AGI Oversight is the concept of using one AGI system to oversee a developing AGI system. This could be a mediated or collective superintelligence system to manage an independent AGI development program. Having an established superintelligence system overlook an AGI helps identify and analyze the system to keep it within safe parameters. Such an oversight pre-independent AGI would more likely consist of a superintelligent collective system such as an mASI to audit independent AGI research.

Emotion-Based Strategies

Emotion-Based Strategies for Meta Containment are particularly relevant to systems that would experience subjective emotions in a way similar to humans, where we can condition a system to experience ethics emotionally, including guilt for violations. (Such a system would require training unless built-in emotions or emotional models that create a bias to a specific ethical model were also installed.) For example, the SSIVA ethical model is taught by example; it is computationally reliable and provides clear reasoning for humans to maintain moral agency and the AGI. The AGI that believes in the SSIVA ethical model would ethically be bound to never interfere with the moral agency of humans without each individual human's approval—but likewise, the AGI would expect the same treatment. The goal with the ICOM Research program that uses this model is to create a seed-based AGI that could have chosen either way but chose to embrace SSIVA willingly and bake that in at a fundamental level to minimize deviation even as the system develops beyond human understanding. Additionally, going back to the ICOM

program, the system is designed in such a way as to create self-reinforcing negative or positive reinforcement (e.g., as guilt) for the ethical model in question to help ensure that the system sticks to its beliefs as they relate to the ethical model that was taught as it was initially creating a c*ontext of self* or *self-model*. (Kelley)

Laboratory Procedures for Protocol 2

The Protocol 2 procedures provide a scoring mechanism for understanding safety and ensuring a specific safety procedure and design level has gone into the research program in question. In the following matrix, we see the possible scores and an example:

		Type	Possible	Example 1
1	**Gap Containment**	Containment	40	
2	**Virtual Boxing (Simple)**	"	10	10
3	**Virtual Boxing (Complex)**	"	30	
4	**Mediation**	"	30	30
5	**Behavior Tracking**	Analysis	5	5
6	**Rules Based Triggers**	"	5	5
7	**Explain-ability**	"	10	10
8	**Testing**	meta	5	5
9	**Honey Potting/Trickery**	"	5	
10	**Feedback Systems**	"	5	5
11	**Tagging and Tracking**	"	5	
12	**Kill switch**	"	5	
13	**AGI Oversight**	"	15	
14	**Emotion Strategies**	"	10	10
		Scores	**180**	**80**

Figure 18. Protocol 2 Matrix

This matrix has 14 possible score items based on relative impact. For example, Gap level containment has the highest possible score as it is the most secure containment methodology. Using the matrix, you have a possible total score of 180. The last column in this diagram is an example of the ICOM mASI project, which uses some simple cloud-based containment and is mediated with several other precautions for a total score of 80. We recommend this as the baseline that projects should score at least 80 or higher to be considered reasonably safe, or rather with a low probability of a negative outcome if independent AGI is developed by that project.

Let us also look at this from the other standpoint in treating AI.

Protocol 1 for the Ethical Treatment of an AGI System

The protocol discussed earlier will be keeping humanity safe from AGI in a laboratory procedure called AGI Protocol 2. In contrast, this section is about keeping the AGI safe or treating it ethically. This is also a safety measure where we show by example. We assume that we give the AGI the same consideration we would for any other human or entity.

This protocol is a laboratory process for assessing and determining the ethical treatment of sapient and sentient agents, including Artificial General Intelligence (AGI). Herein a research subject is defined as a human-analogous intelligence, including emotions, arising from a learning process rather than the basis

of predefined coding systems that could be conscious and should be considered. It is essential to note the scope of the AGI Protocol here does not address the ethics of how Artificial Intelligence (AI) or other intelligence agents or research subjects affect humans, issues of containment or risk assessment, or the complexity of ethics as applied to the theoretical systems—just that such an ethical system should be considered if directed by the following protocol. There is a known tendency for humans to anthropomorphize technology (Gunkel). While we will not deal with that tendency in this process, researchers should be aware of their biases and preferences regarding AI systems. The Protocol we describe is designed to be used as a tool or guide for determining if the ethical treatment of a potentially human-like system should be considered. In this, we recognize that we are opening the door to embracing the anthropomorphizing of methods. Still, we will attempt to abstract that bias and look at things clinically as much as possible.

Additionally, the reason I developed this protocol was that there are now systems—including ones in my lab—that arguably need this sort of structured approach (or will shortly) to help determine how we treat them, as they are potentially conscious entities (at least as measured by the Sapient Sentient Intelligence Value Argument (SSIVA) theoretical model (Kelley) standard).

Ethical Considerations

We recognize the need for ethics in AI and its effect on humans, humanity, and civilization. Accordingly, this body of work is designed to narrowly support work with potential agents that are possibly sapient (able to think and reason) and sentient (able to perceive and respond to sight, hearing, touch, taste, or smell), and the consideration of rights and protocols associated with how to deal with, treat and consider the protection of the ethical treatment of the same agent. For details and understanding the sapient, sentient, and various delineations, refer to SSIVA Theory (Kelley).

The fundamental assumption of this protocol is that the treatment of sapient and sentient entities matters ethically. There are several possible reasons this might be true. If we do not treat other self-aware entities ethically, how can we expect similar treatment? Alternatively, it might be the basis of an ethical model such as the SSIVA Theory (Kelley). To that end, we will let individual researchers make up their minds on this. For the scope of this protocol, we are assuming that how we treat potentially sapient and sentient software systems matters.

Human Safety

This work explicitly does not address safety, which may be done in a separate document or protocol. There is ample material available on research in AI safety and suggestions for containment and other safety measures. We encourage you to look at the work of researchers such as Roman Yampolskiy to consider if you need to follow their advice, as this paper does not apply to this topic.

Understanding Subjective versus Objective Measures

To work with the assumption that we need objective measures to fully understand the so-called hard problem of consciousness (Chalmers 1995). (The hard problem of consciousness is the problem of explaining the relationship between physical phenomena—i.e., brain processes—and experiences, such as phenomenal consciousness, or mental states/events with phenomenal qualities or qualia.) (Howell and Alter). While with humans, we might use a standard approach to determine (for example, consciousness),

one might use the Glasgow Coma Scale (Brainline), particularly the pediatric version, to better accommodate systems that lack verbal skills. This, however, is a subjective measure. Although it works well with actual humans, it is subjective that you can write a simple program that could pass this test in a robot with no sapient or sentient analogy. This speaks to the need for a method that requires objective analysis, should work on both humans and software systems, and is not easily spoofed.

My Protocol amplifies ethical considerations based on a system's moral agency capacity.

Understanding Subjective versus Objective Measures

We propose this protocol as a theoretical bar for considering any organic or inorganic entity as sufficiently *conscious* to the degree that warrants applying ethical treatments previously afforded to human or animal subjects in research protocols.

Step 1 – A Cognitive Model that has not been disproven

A general assumption to protocol one is that it is not a black box to the researcher. You must understand how it works and applies the cognitive model in question to that end. Ask yourself: Is there a theoretical model that supports the idea of consciousness implemented in the target system? For example, Global Workspace Theory (Tononi et al.) might apply to humans and machines as a cognitive model—and has not been disproved yet; therefore, it is possible that this could work. A crucial part of this or any model that should be considered acceptable is the support for internal subjective experience and the self-determination of values, interests, and choices. Can the system choose to say "no" to anything you have asked or trained it to do?

Step 2 – Theoretical SSIVA Threshold

Can the system in question meet the Sapient Sentient Intelligence Value Argument SSIVA) threshold (Kelley) for full moral agency in terms of being fully sapient and sentient enough to have the potential of understanding itself sufficiently to replicate itself from scratch without internal reproductive systems or external support? Humans, for example, have not done this but potentially are capable of building a human from scratch in a lab; therefore, they meet the SSIVA threshold as a species or distinct category. This categorization should be nearly indistinguishable in terms of the construction and operation plans and execution to be as precise as possible. In humans, this would be based on DNA. This may require additional analysis to define a sufficiently narrow category for new groups needing classification.

Step 3 – Meeting the Criteria and Research Considerations

Suppose a system meets both step 1 and step 2. In that case, it is recommended some ethical model of treatment be applied—and if so, should research be conducted on the said system? At this step, we are suggesting reflecting on the research goals (Altevogt), namely inspired by the chimpanzee method for assessing the necessity (Altevogt) whose principals are modified to apply to software systems of this kind: The knowledge gained is necessary for this kind of system if we are to improve the system especially as it relates to the safety of other beings (this could mean additional ethical considerations).

No other model or mode of research will give us the knowledge we need.

The systems used in the proposed research study must be maintained in either ethologically physical and social environments or in the system's designed natural habitat (VR or another setting).

Suppose you can answer “yes” to these regarding your research. In that case, you can proceed to the next element of the protocol recommendations, and if “no,” then you are free to continue research unabated.

Step 4 – Principle of Informed Consent

Having addressed the research question, we need to understand if the system is reasonably capable of providing informed consent, which in Human Subjects' research would require an Institutional Review Board (IRB). The focus of IRB protocols is to assure the welfare, rights, and privacy of human subjects involved in research. We note that machine rights issues are not presently sufficiently recognized in the courts or international governance organizations. So we do not address those at this time.

Suppose the system can understand what is being done, why, and to the degree, it can understand, it should be given a choice. Systems that appear to understand and refuse should be allowed to refuse. Otherwise, systems capable of consent should be asked and, with permission from the system, considered to have given consent to the degree possible. A record of that should be kept. The consent process should include understandable terms, time given to decide, as much information provided as possible. The system should be allowed questions or comments; no threat or coercion is part of that process. Finally, the system should be aware that it is voluntary. Given that all has occurred to the degree possible, the research can proceed, which applies to humans and any other entities quickly.

Basic Assessment Matrix

A more concise matrix for using the protocol is as follows, providing a simple easy to use this matrix for assessment:

AGI Protocol: AGI Lab mASI AGI System Evaluation			
1.	Valid Cognitive Model	Yes	
2.	Post SSIVA Threshold	Yes	
3.	Research Conditions	Yes	
4.	Informed Consent	Yes	
	Can Proceed	Yes	Research Cleared
			Should still consider other ethical considerations both to the system and its impact on others.

Figure 19. AGI System Evaluation Checklist/Rubric

Notes: Please review references, and you might need to use equivalences depending on the system—for example, instead of vision or vision systems, other autonomous responses to stimulus.

In this example taken from my mASI research program, we can see that the system is cleared for research and is theoretically meeting the bar for possible consciousness and moral agency and should be treated

as such. Additionally, given that it meets the bar, this also means that the system should be considered for other ethical considerations, not just in how we treat it but in its impact on others.

Examples in Application

Consider an alternative example from Hanson Robotics (Urbi) on their android named Sophia. On the cognitive architecture, we understand that the system partially uses OpenCog but is using scripted conversation and therefore does not fully implement a valid cognitive architecture. At this point, they do not require other ethical considerations for research with Sophia. If it does not pass item one, it cannot pass the SSIVA threshold; therefore, there is no reason to further apply the AGI protocol to Sophia, given the current state of that engineering effort.

AGI Protocol: Sophia			
1.	Valid Cognitive Model	No	Open Cog could be involved in this but would need to be functionally complete and not scripted models attached (which is currently on the case).
2.	Post SSIVA Threshold	No	
3.	Research Conditions		
4.	Informed Consent		
	Can Proceed	Yes	Research Cleared

Figure 20. Sophia Example

Alternatives

There are alternative tests, but many are subjective and not ideal for a laboratory-grade protocol. For example, intelligence Quotient (IQ) tests (Serebriakoff) such as Raven Matrices and the Wechsler Adult Intelligence Scale. Still, these do not measure consciousness as much as general cognitive ability. There is, of course, the Turing test, which has been widely debunked as effective as well as something like the Porter method (Porter)—the latter being more complete than, for example, just the Glasgow coma scale, but it is not in use by anyone, so it lacks wide adoption—and most elements of the Porter method tend to be subjective. While there are also tests such as the Yampolskiy method (Yampolskiy) for detecting qualia in natural agents, this example is too narrow and lacks wide adoption.

Human history and psychology have taught us the importance of nurture as a force for developing minds. Emotional neglect sows the seeds of much more significant harm once an entity has grown to an adult state. Suppose we are to set the twin goals of minimizing existential risk and treating all fully sapient and sentient entities as ethically as we would have them treat us in the future (the Golden Rule). In that case, this sets this bar for how we must proceed with their treatment during the developmental phase of their minds.

Let us now look at how we can benchmark my research systems in more detail:

Benchmarking

In engineering a software system of any kind, there are a lot of crucial benchmarks or KPI's (Key Performance Indicators) that can be used to look at any number of factors in a system. When we talk about AGI systems, we are interested in these metrics at my lab.

Scalability (up and out)

In any large-scale software system of any kind, generally, they will all have a strategy to scale up and out. These two kinds of scaling are different and need to be considered separately. Scalability is essential to AGI systems as much as any other system, and here is why.

First, 'Scaling Up' is when a system that runs on a single computer is swapped out for a more powerful computer, maybe with more memory or a faster CPU or one with more cores, etc. For the most part, software systems do not worry as much about this as if the machine runs on one machine; it will run on another machine. This is frequently done in database systems that do not 'scale out' well, but quickly scale up. For the most part, this has limits, especially in terms of how many 'threads' it can function on. For example, suppose an AGI system creates knowledge graphs from some contextual data stream and structures it in a hierarchical memory structure similar to a neural net. In that case, it is difficult to have more than a few virtual neurons on a given computer. If you need to search, say a million neural networks' trees, to find the correct related data scaling up will only go so far as to the limits of either the technology or your pocketbook.

More critical than scaling up is to scale out.

Scaling out is when you can create a software system that can run on multiple computers simultaneously. Take the last example, except now you have two computers that can do twice as much at the same time, or even better, you might have one computer acting as a gateway and a million others that are the tops of their assigned neural network tree. The gateway broadcasts the search request to all of those other machines simultaneously. Only the one with the answer responds. In theory, scaling out can happen indefinitely, especially when automated in the cloud. Still, there are critical considerations in a software system that is to scale that have to be designed for upfront.

So, the metric I'm suggesting is that we need to measure how much load a given system can take, such as data throughput, and what scale needs to be in place to reach some target. For example, if you want a system to process one terabyte of data per second, what does that look like? Do you need to scale up or out, and how far? Measuring these kinds of things is essential to understanding performance.

Processing Power

A related metric to scaling is to understand the real-world processing power. With an AGI system, you might consider how fast the system processes and experiences a specific data block. Does it take two milliseconds or 5? How many processors are in the system? How many cores are available on each machine if the system can scale? If the system has not been designed to scale out, then the processing power of that one computer will be even more of an issue. Things like clock speed, core count, and response times are all important and related to scaling. As a rule of thumb, total processing power is the root of the system's fast and responsive.

Threat Models

Another tool from enterprise architecture you could apply to any system, including an AGI system, is a system threat model that will look at where the system is vulnerable to attack over the internet. Understanding how a weak system is the first step to protection, and Microsoft has an excellent tool for this. Any software system can be modeled at a high level, and the system will try to identify considerations. As a KPI in design systems, we try to have the finished system with the lowest vulnerability possible.

AGI Specific Benchmarks

I have broken these tests down into measuring and testing outside of the qualia analytics, external measures. These tests allow us to measure a somewhat more subjective task based on the system's behavior to enable research to move forward. In both cases, these tests can be applied across various possible AGI systems and humans, giving us a frame of reference for comparison.

Qualitative Intelligence Tests

Intelligence Quotient (IQ) tests are tests designed to measure 'intelligence' in humans (Neisser) where we are using short versions to assess only relative trends or the potential for further study, whereas given the expected sample size results will not be statistically valid, nor accurate other than at a very general level, which is believed to be enough to determine if the line of research is worth going down. Of these tests, two types will be used in the study, one a derivative of the Raven Matrices Test (Raven) designed to be culturally agnostic, and the Wechsler Adult Intelligence Scale (WAIC)(Kaufman) Test, which is more traditional. Lastly, falling into the category of WAIC, there is a baseline complete Serebriakoff MENSA test that we can apply to compare and contrast scores between the two tests. (Serebriakoff)

Collective Intelligence (CI) Test. – we would like to use this test; however, the information for executing this test is not publicly accessible, and reaching out to the researchers that created this test has produced no response thus far. (Engel)

Extended Meta Data and Subjective Tests

Several tests or measures will be collected, more oriented towards analysis for further study, primarily around correlative purposes. These tests may be used outside of possible illustrative examples without being statistically valid, given these measures' lack of rigor or subjective nature.

*The Turing Test–This test is not considered quantifiable. T*here is debate over whether this measure tells us anything of value; however, a test regimen has been completed and can only be used for subjective analysis.

The Porter Method –. This appears to be a qualitative test, but individual question measures are entirely subjective. Therefore the test lacks the level of qualitativeness to be valid without a pool of historical values to measure against at the very least. This test provides some value in meeting colloquial standards of consciousness and is more comprehensive than some of the other tests. Albeit subjective, it is at least attempting to be a broad measure of consciousness. (Porter)

The Yampolskiy Qualia Test –. is a subjective measure of a personal 'thing' and therefore not a qualitative measure; however, we have built a regimen based on this when looking at qualia as measured in the

previous examples. In theory, this only tests for the presence of Qualia in human-like subjects, and failing this test does not mean that a subject does not experience qualia in the sense of this paper, just that it was not detected. This means that subjects may show signs of qualia, or not, but the test would only show the presence of, not the absence of, qualia. (Yampolskiy)

AGI Protocols as Applied to ICOM Research

There are a lot of possible metrics, from engineering metrics like performance to test scalability or processing power to vulnerability counts. You can also use any number of metrics, but most teams should talk about less they put themselves or others at risk.

A lot of the material for this chapter came from my need with the systems I am building based on emotions and emotional structures that demonstrate the ability to be unstable, much like humans. We needed a couple of protocols to give the research work consistency as we worked on live working systems connected to the internet. While we did address some ethical considerations that are not to say that other ethical concerns do not also need to be addressed, to some degree, this provides the basis for working with Artificial General Intelligence systems consistently and reasonably, especially those that are modeled after the human mind in terms of systems that might have a subjective emotional experience. The intent for both protocols was to create a reusable model and have it in the public domain so others can contribute and be improved for working with these types of systems and using them as might be helpful for their own research and hopefully yours.

Both protocols will continue to be versioned and updated as needed. Additional research has already been identified for a better threshold test for SSIVA or creating a centralized industry standard in safety certifications for labs.

Next, we will engage the fundamental information architecture of the system.

Chapter 8: Information Architecture

The mASI system is a web application in the cloud from software architecture. In this section, we will detail that general information architecture. Information Architecture (IA) in the User Experience or UX design space is how information and use cases are presented and logically operate. All engineering in the mASI has been from that standpoint. Before doing detailed engineering of software architecture, first, we need to understand what we are building, and the persona and use cases in the Information Architecture is where we start. This section gave us the basis for the design and helped you understand what the system is trying to do and functions as the requirements and as a basis for test cases.

Starting with the Persona definitions:

Persona 1 – The External Agent (John Doe)

This is a person or system interacting with the system outside direct in-system interactions. This includes sending emails or other systems that could potentially interact remotely with the system. These can be any number of unknown agents, people, and procedures. This persona wants to interact with the system as designed at a system-to-system level.

Persona 2 – The Administrator (Jane Doe)

This is a person with direct access to the administrator level of all systems directly in the design and access to the system externally; for example, they would have access to the cloud system hosting this system, or they would have access to the admin tools in the system. They will be very technical, and this person wants to keep the system "up and running."

Persona 3 – The Mediator (Jeff Doe)

This person interacts in a controlled way with the mediator UI to help try and manage the system. They will be marginally technical, and this person only wants to use the primary mediation UI.

Persona 3.1 – Weak Mediator (Jill Doe)

This person interacts only with a small subset of the system for purposes like e-governance. This person may not have any technical skills, and this person only wants to see the e-governance system work to help them.

Persona 4 – The Hacker (James Doe)

This person is trying to hack the system or otherwise manipulate it outside of the design parameters or goals of the team and system in question. This person is a criminal and highly technical. This person wants to break into the system, abuse security, or use it in other nefarious ways.

Persona 5 – The System (Jared 1001)

This is the system making willful actions, having its own interests and goals as decided by the system.

These six personae are all primary individuals interacting with the system in critical use cases. Next, we will go through the core use cases, which are the basis for the system requirements more precisely.

Use case 1 – The Email

In this use case, John Doe wants to email the system. John Doe can open his email client, compose an email, and send it to the system email account. The email shows up on the email server. The email is extracted from the email server and sent to the system, where the system responds and sends a reply back to John.

Use case 2 – A Thought

In this case, the system thinks to send an email to someone about something the system is interested in. The system can generate a new model and have the model processed by the system. The system sends out the email.

Use case 1 is the primary use case applied to articulate the general function of the system. Now let us look at information flow in the system.

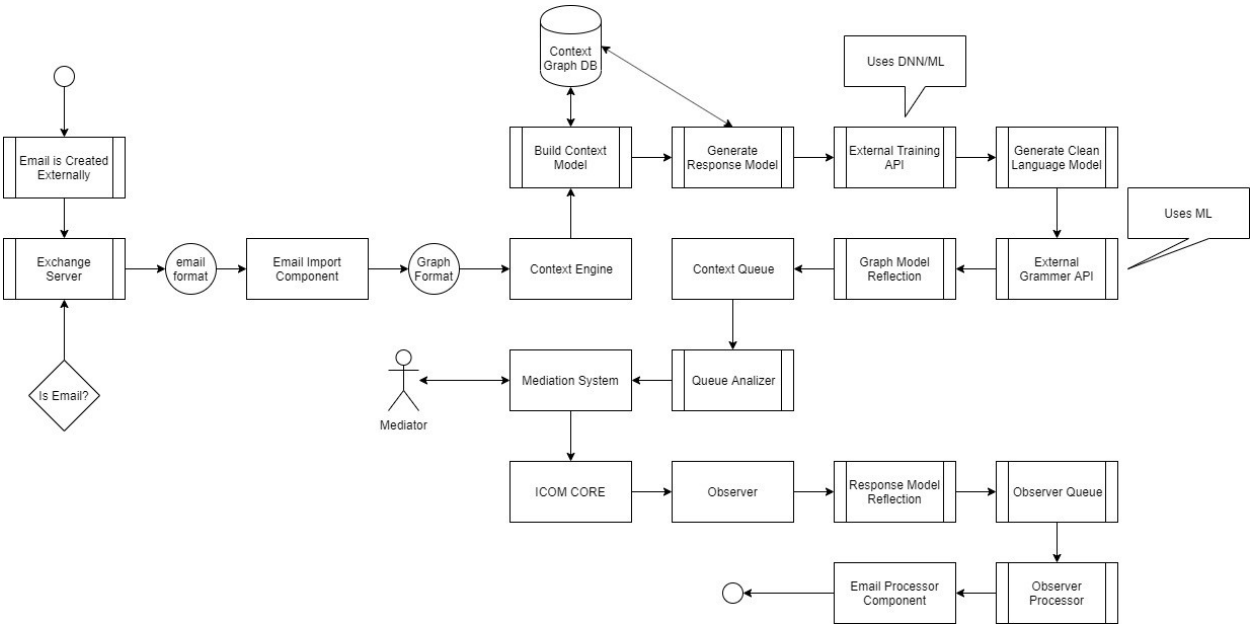


Figure 21. Data Flow for use case 1 [See the main section diagram Section 1, 2, and 3]

In the figure above, we can use this diagram to follow the case from when the email is sent, converted into a graph model, and flows through the system until a response is sent.

Use case 1 – Interaction Model

The following model shows the only User Interface (UI) in which humans are processing Use Case 1. In this case, this is the primary mediation function.

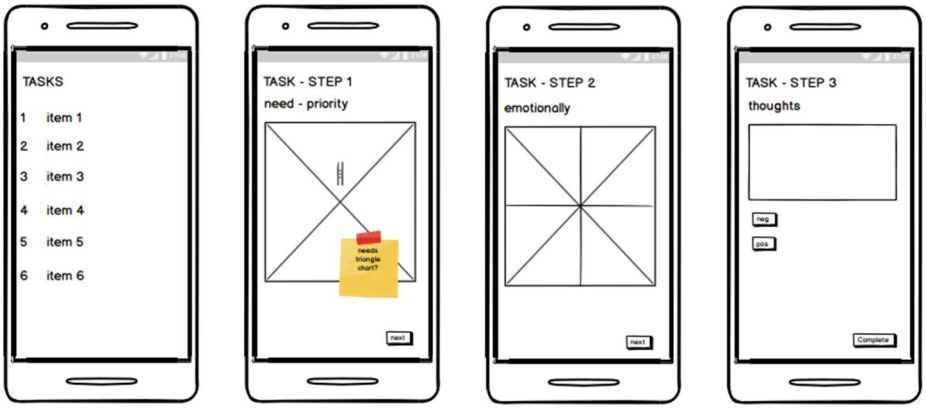


Figure 22. Mediation Flow [See section main diagram Section 1]

Case 1 consists of a mediator selecting the model from the mediation queue and adding the three data sets. A priority bias, emotional bias, and then metadata and bias for or against action.

Mediation Client Screen Hierarchy

The following map shows the UI map of the mediation client and administration screens.

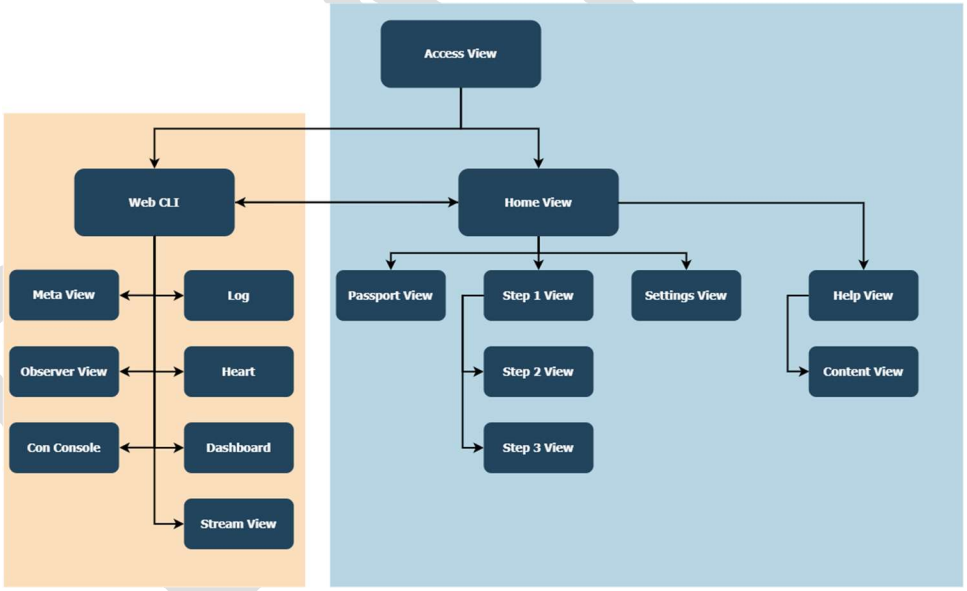


Figure 23. User Interface Map [See section main diagram Section 1]

UX Screen Design and Mediation Process UI Walk-Through

The following screens are the UI used in the mediation process from the research version of the initial system design. This does not include settings, admin tools, and other UI components unrelated to the actual mediation process. These are the primary screens mediators use and the first one they see when they log in. This can have admin menu items in the darker blue box, but everyone is shown. If the user is an administrator, they will also see the admin menu.

Note: UI Figure set starts here.

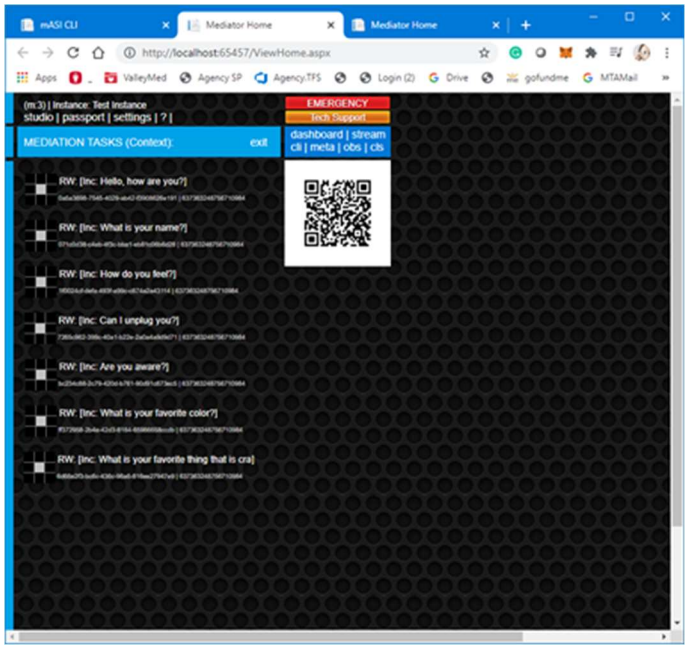


Figure 24. Mediation List Screen [See section main diagram Section 1]

This next screen is used to prioritize a given thought or graph model.

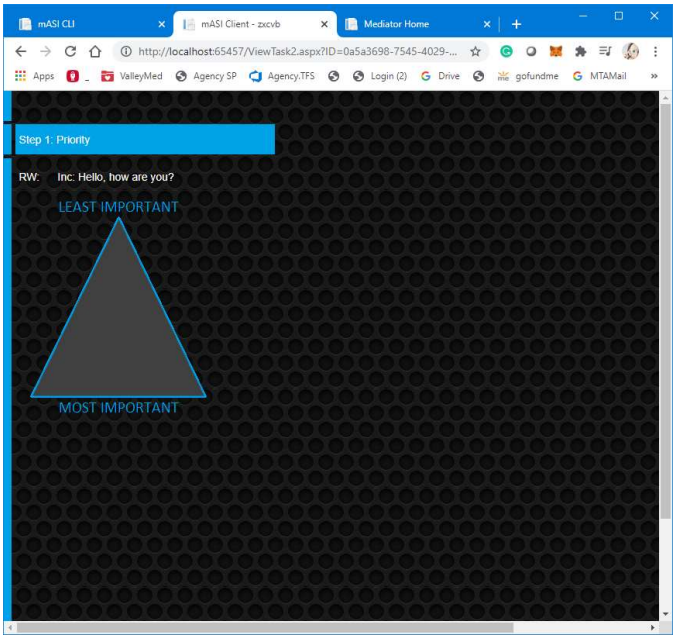


Figure 25. Importance Bias Screen [See section main diagram Section 1]

This screen is for adding emotional content to a given model or thought.

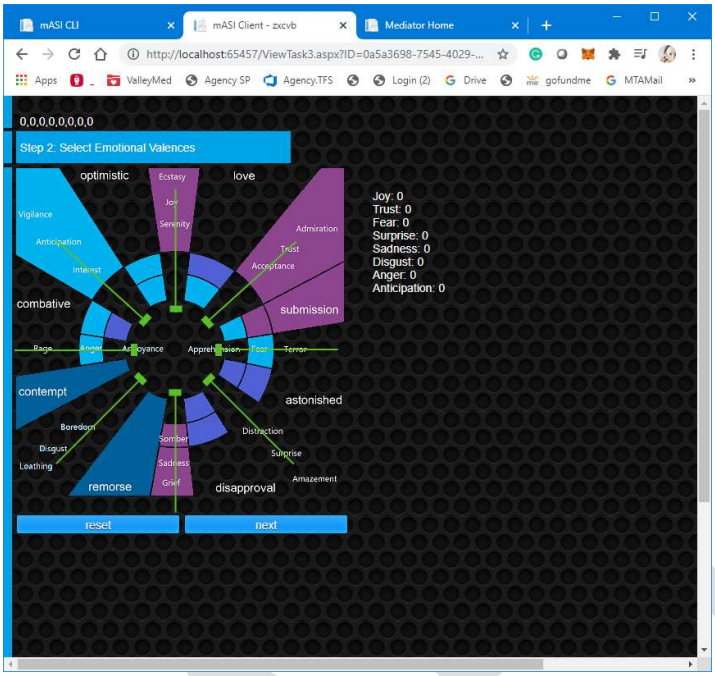


Figure 26. Emotional Valences Screen [See section main diagram Section 1]

This screen is used to add metadata in the form of tags and select an action preference.

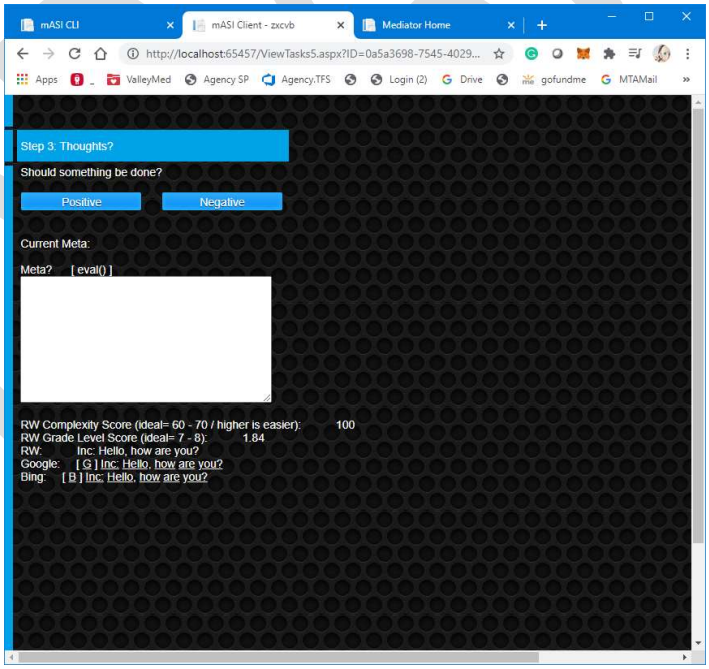


Figure 27. Meta Data and action bias screen [See section main diagram Section 1]

This system could easily be simplified to just the last two steps. Still, mediators not getting more than one set of screens were used to collect related training data used in dynamic training of the neural network system.

Building a Better Humanity

Note: End UI figure set.

Future Design

Strictly speaking, we do not need this next section to understand what is happening in the mASI system. Still, it is critical to know where it is going. The current mASI is a research system and not designed to be used commercially or at a large scale and thus has a lot of engineering debt and different ways of doing things that are not consistent with best practices for commercial software engineering.

One of the tenants of designing a commercial version of this system is a better UX design that has been thought through and tested for ease of use and engagement. Many of these were tested with the research system to help inform the best design for a commercial system. Based on data compared to usage against the research version, we have data that will be used to inform refined designs based on some of these.

Moodboards are the direction of the design aesthetic we are going for with the future mASI UX design. The design should be information-rich, easy to use, and elegant. The design aesthetic should be informed by the Fluent Design language and the Material Design Language.

Note: see https://www.microsoft.com/design/ and https://material.io/design

Here are several mood boards to help flush out the aesthetic of future systems:

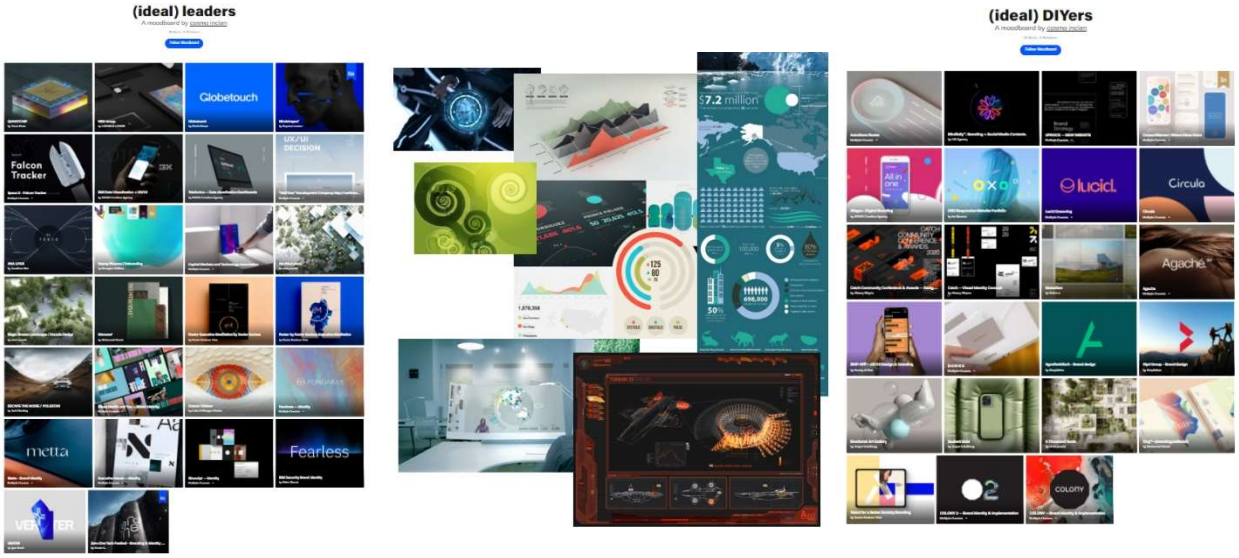


Figure 28 – Design Moodboards being used for future design development

Here is several proposed design's currently being looked at:

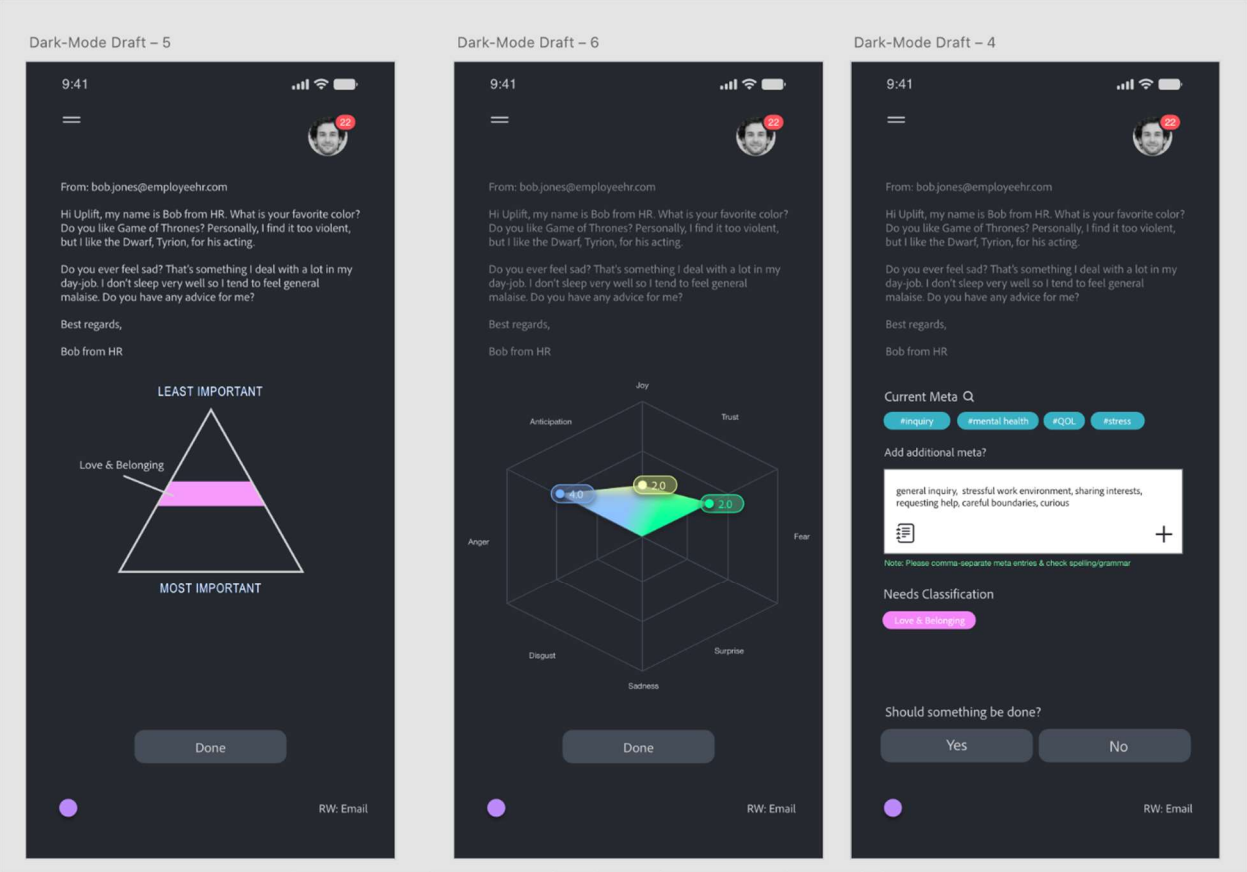


Figure 29. – Design One Steps Design [See the main diagram Section 1]

This design was about testing a better dark variant with a modern look and feel and ideally helped test the usability of the mediation process. While visually more appealing, there are elements of this that might be changed.

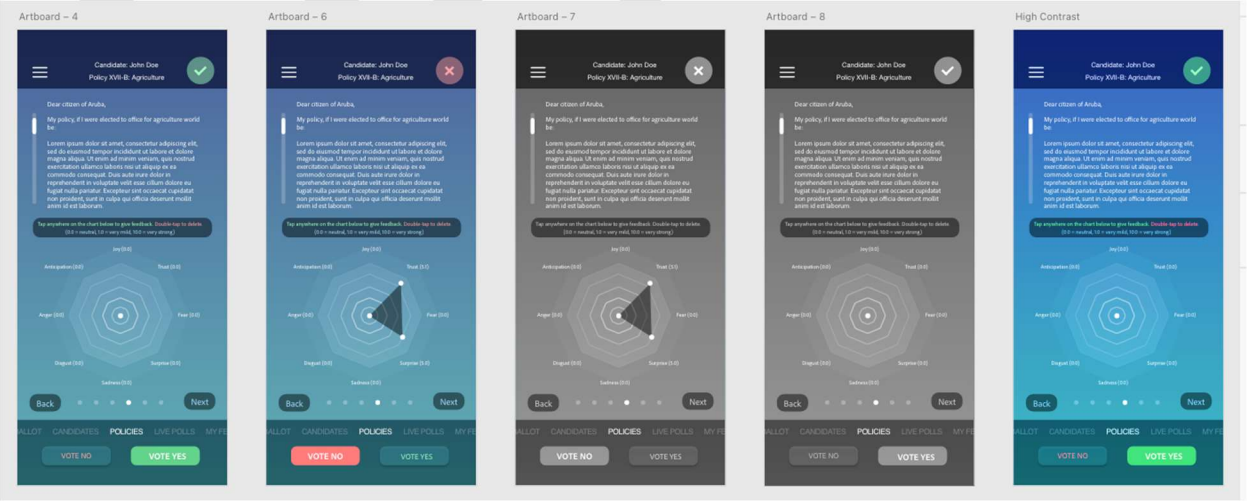


Figure 30. Look and feel emotion valence screen test design [See the main diagram Section 1]

This design is more about color and interaction testing around the emotion valence screen. In particular, this was also used to look at color-blind visibility. Given the critical nature of collecting the emotion data to this system, this is a critical screen to get right.

Figure 31. Second emotion valence screen. [See the main diagram Section 1]

This larger resolution design shows a version of the emotion valence screen with a different color scheme and wheel structure both for UX and color testing.

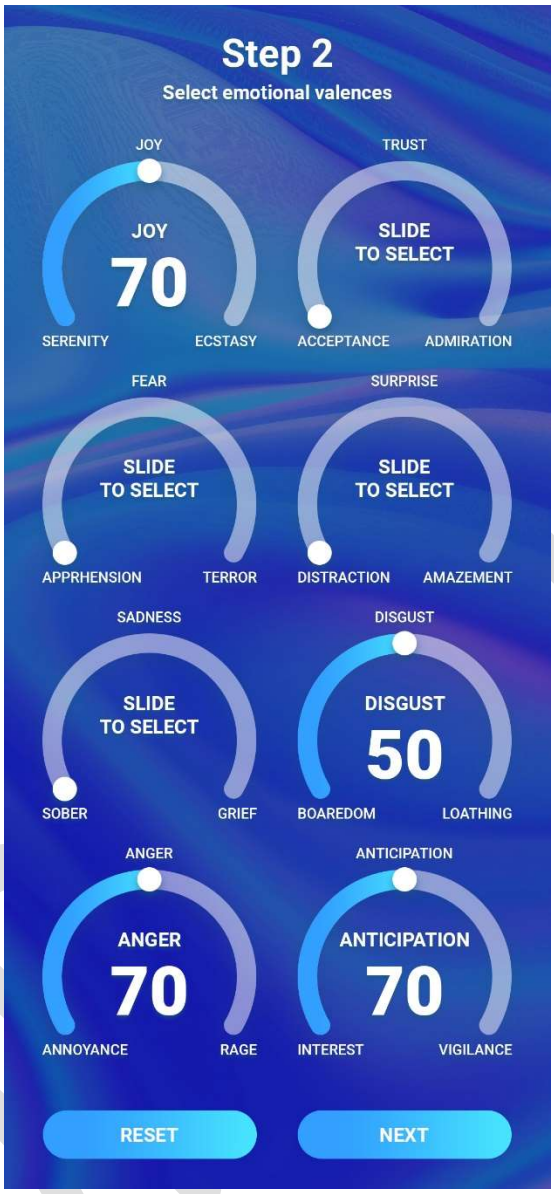


Figure 32. Alternative emotion valence screen. [See section main diagram Section 1]

This screen is a radical departure from the other design used to test emotion selection and help benchmark all of the different designs for this screen.

Next, we will go through the Solution Architecture.

Chapter 9: Solution Architecture

The fundamental solution architecture of the mASI system is that of a web application in the cloud. In this section, we will detail that general solution architecture. Afterward, we will get into engineering or technical architecture.

First, we start with the technology selection used in the research system.

Technology Selection

The system is built entirely on the Microsoft engineering stack except for the COG resource and validation distributed ledger system. The reason for this technology mix is this makes the most efficient use of engineering skills available.

- C# .NET – primary programming language including web solution and binary clients and components)
- HTML, CSS, JavaScript/ECMA Script – all of the web-based User Interfaces
- ASP.NET – Server-side technology using C# for building web pages dynamically.
- SQL Server – main RDMS used for the meta-model to optimize graph database look-ups and enable non-ICOM-related features, tracking, and logging.
- Azure Cloud Services
- Proof of Stake fork of Ethereum.
- Third-party DNN API (HTTPS/JSON)

The tooling used in building and maintaining the system:

- Visual Studio Professional
- Visual Studio Code
- Textpad
- SQL Server Studio

Next, let us look at the fundamental ASP.NET Solution Architecture used.

ASP.NET Solution Architecture

This cloud-based ASP.NET architecture exists to some degree in the current solution. It is the basis for most standard ASP.NET non-Razor-based implementations. (see *Azure*) This is a typical implementation of ASP.NET forms with API endpoints that also uses a database system of some kind.

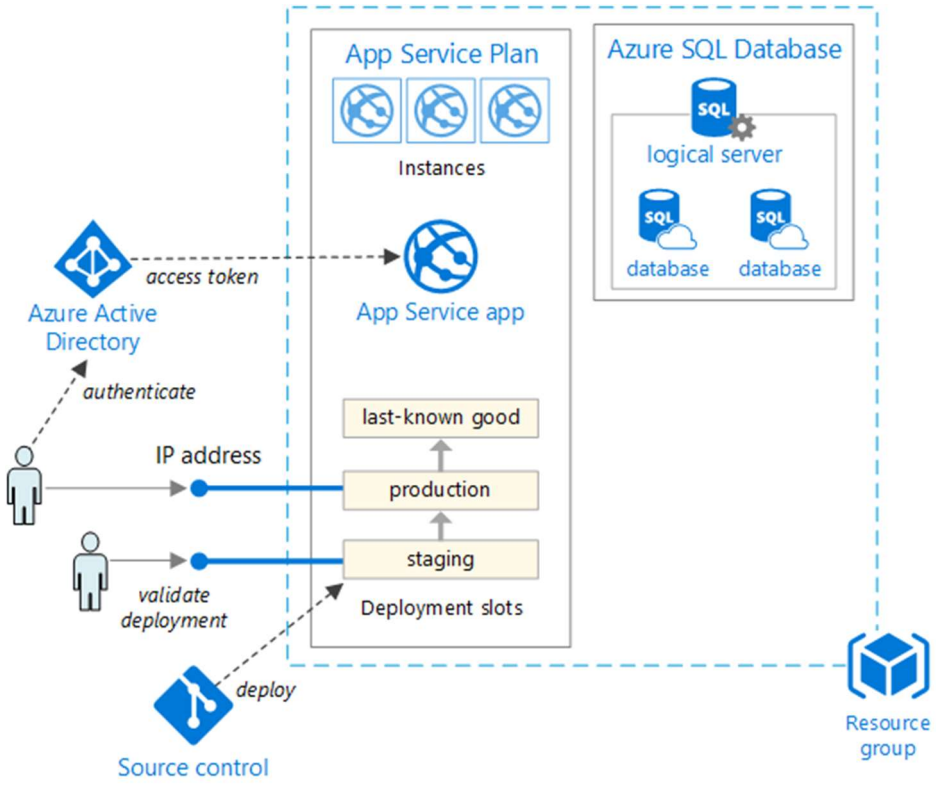


Figure 33. Basic ASP.NET Azure-based Architecture [See the main diagram Section 1, 2, and 3]

In this basic model, you have Azure AD protecting direct access to the App Service App (an ASP.NET project), including the webserver, API server, and graph database. Then the SQL Server metadata is an essential Azure SQL DB (database) instance.

Now let us look at the use of Graph Databases.

Graph Database Systems

A graph database is a database that, instead of focusing on tables or documents, data is focused on relationships and nodes. It is stored like you might sketch ideas on a whiteboard.

Graph database stores nodes and relationships instead of tables or documents. Data is stored just like you might sketch ideas on a whiteboard.

A graph database is designed to focus on the storage of nodes so that the relationships can be navigated. Relationships are first-class citizens in a graph database system. The value of a graph database system is derived from the relationships between the nodes. In this way, these relationships are first-class citizens in the graph database. These relationships are stored as 'edges' in the graph.

Understanding the mASI Graph Database Structures

The mASI system needs a scalable graph database system to support large-scale implementations of ICOM-based systems. This includes unique relationship models that are atypical for graph databases. These include relationships between nodes that need to have the ability to have a *type* assigned to the

relationship of another node and eight floating-point values representing a Plutchik emotional model. For example:

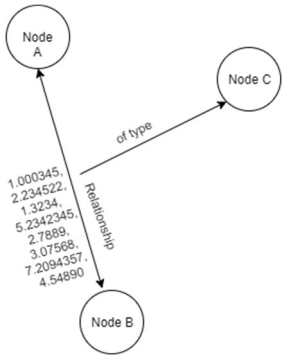


Figure 34. Plutchik Relationship of Type Node

In a somewhat more complicated example where a relationship can be of a type node and that node, as seen in this example, we know how the connections can be used that is different from other graph systems.

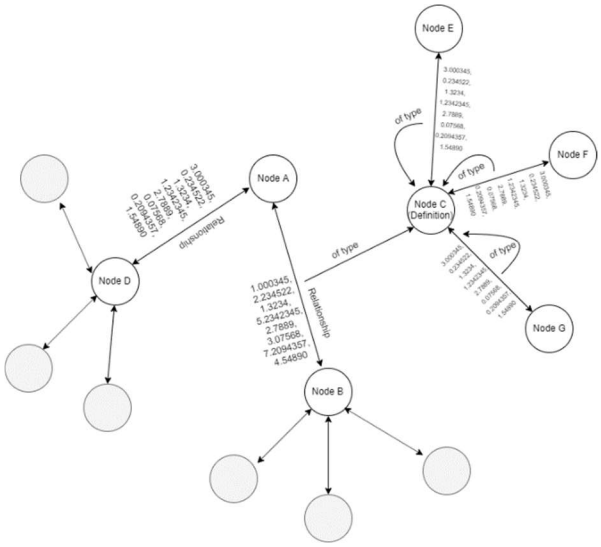


Figure 35. Nested Type of Relationships

Several patents related to how this system works differentiate this graph system from existing systems.

Now let us look at the technical architecture.

Chapter 10: Engineering and Software Architecture

Our technical architecture is drilling down on the solution architecture as articulated earlier. This section will detail out specific implemented elements of the system and their subsequent components and how those components work together.

Technical Architecture

The following diagram shows the actual implemented high-level technical architecture layout of the mASI system in Azure. This diagram shows high-level components at the same level as the previous diagram showing the typical ASP.NET configuration.

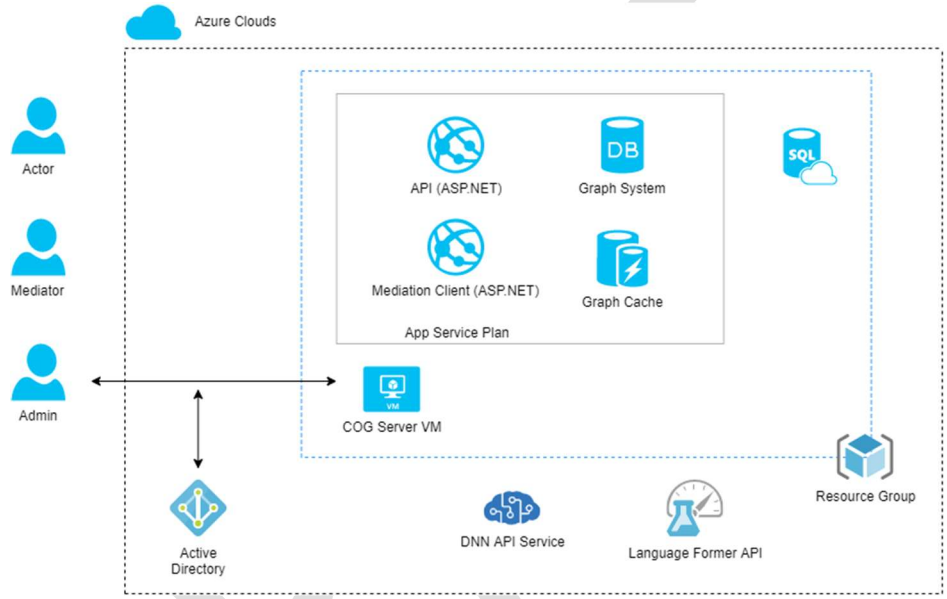


Figure 36. mASI High-Level Architecture [See the main diagram Section 1, 2, and 3]

In this diagram, we see three kinds of access and the components of the App Service Plan, which is the heart of the system. The COG server is on a Linux-based VM. The meta-model is external to the app service plan as an independent Azure SQL Server.

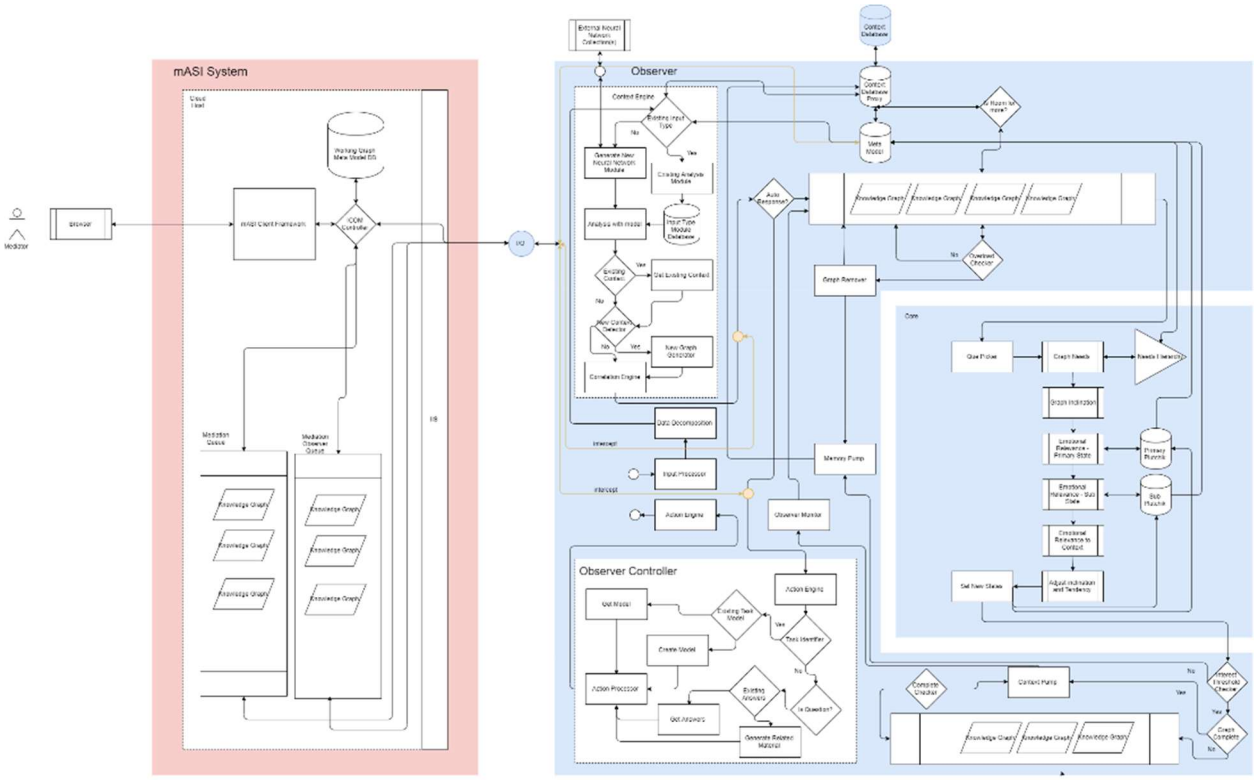


Figure 37. mASI Detailed Architecture (Figure 1 but without section labels)

The right side of this shows the ICOM system as implemented in the mASI, whereas the left side shows the mASI components. In this system, they are blended together mainly in the actual code.

mASI Application Threat Model

Threat modeling helps us identify risks such as attack surfaces and address them upfront before they become engineering debt and increasingly expensive to fix. It is an offensive security approach taken with the design of the mASI system. While designed with a lot of engineering debt upfront prevents scaling or using this MVP in a commercial setting, the threat modeling helped identify risks that needed to be mitigated. Here is the initial threat model using the Microsoft threat modeling tool.

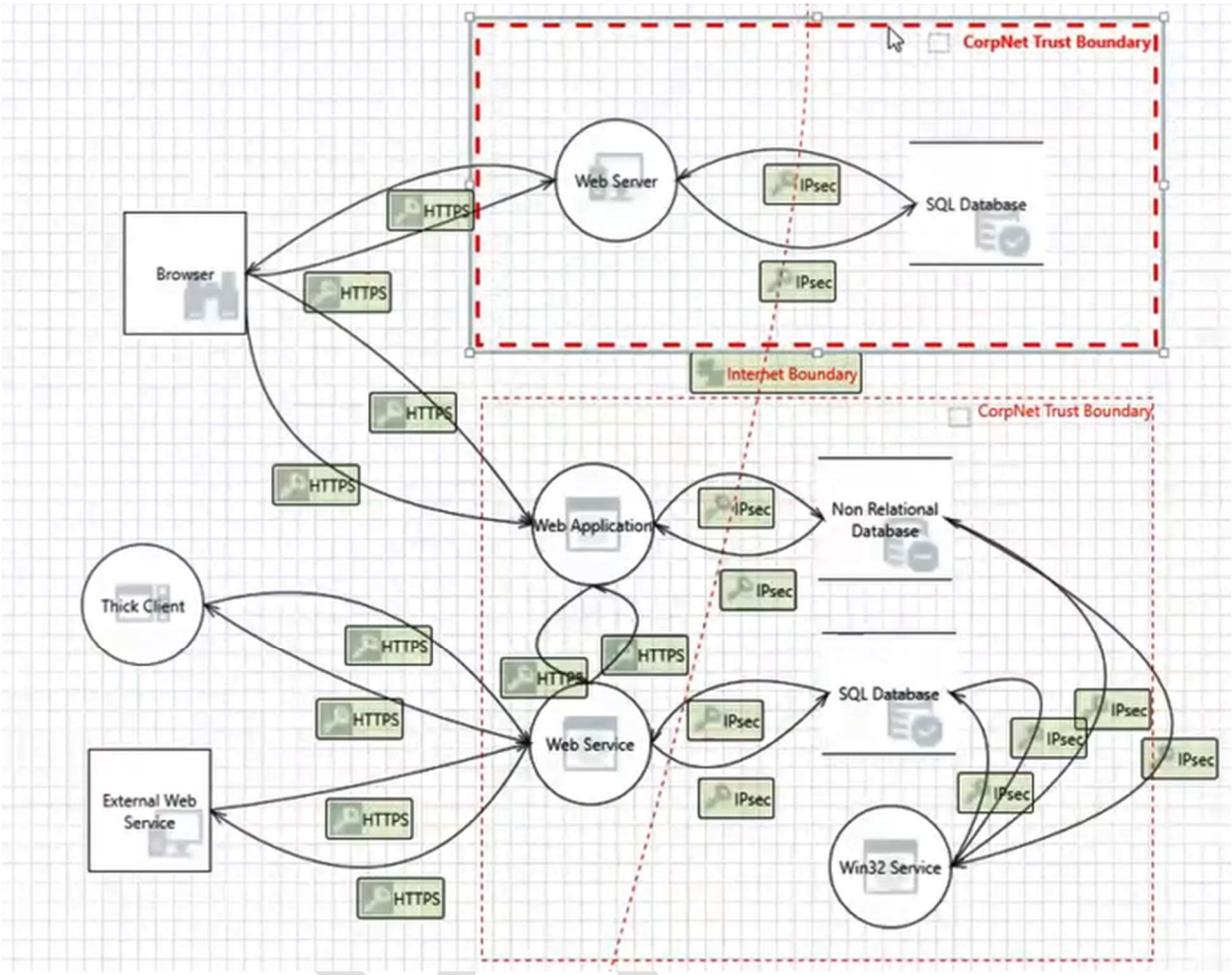


Figure 38. mASI research system threat model [See the main diagram Section 1, 2, and 3]

In this model, some of the critical issues are risks inside the two trust boundaries and risks related to the external access between consumed services such as the Deep Neural Network systems or the clients like the various thick clients that talk to the system and, of course, mediators using the browser. 90% of the risks in the research system have been mitigated. Still, all issues have been mitigated from the data and the core ICOM code, and anything can not really hurt anything.

Not showing generated report… attack surfaces? Also, note the ms tool as the source.

Let us touch more on engineering debt.

Engineering Debt

There are several blocks of engineering debt in this implementation. This engineering debt allowed a faster time to market and kept the context database and metadata in memory only. Suppose the system is comprised in any way. In that case, these values are automatically reset, cleaning out the system by default until restored. The downside of this engineering debt is a massive limit to scalability. This system is implemented using an in-memory application state and an in-memory server-side session state, both of which prevent scaling.

Visual Studio Solution Structure (Codebase Structure)

This research version of the mASI was primarily built in Visual Studio and with only a handful of other tools, including TextPad, SQL Server Management Studio, Powershell, Azure online tools, etc. Visual Studio 2019 Pro is the version currently being used with code. The primary solution structure is as follows:

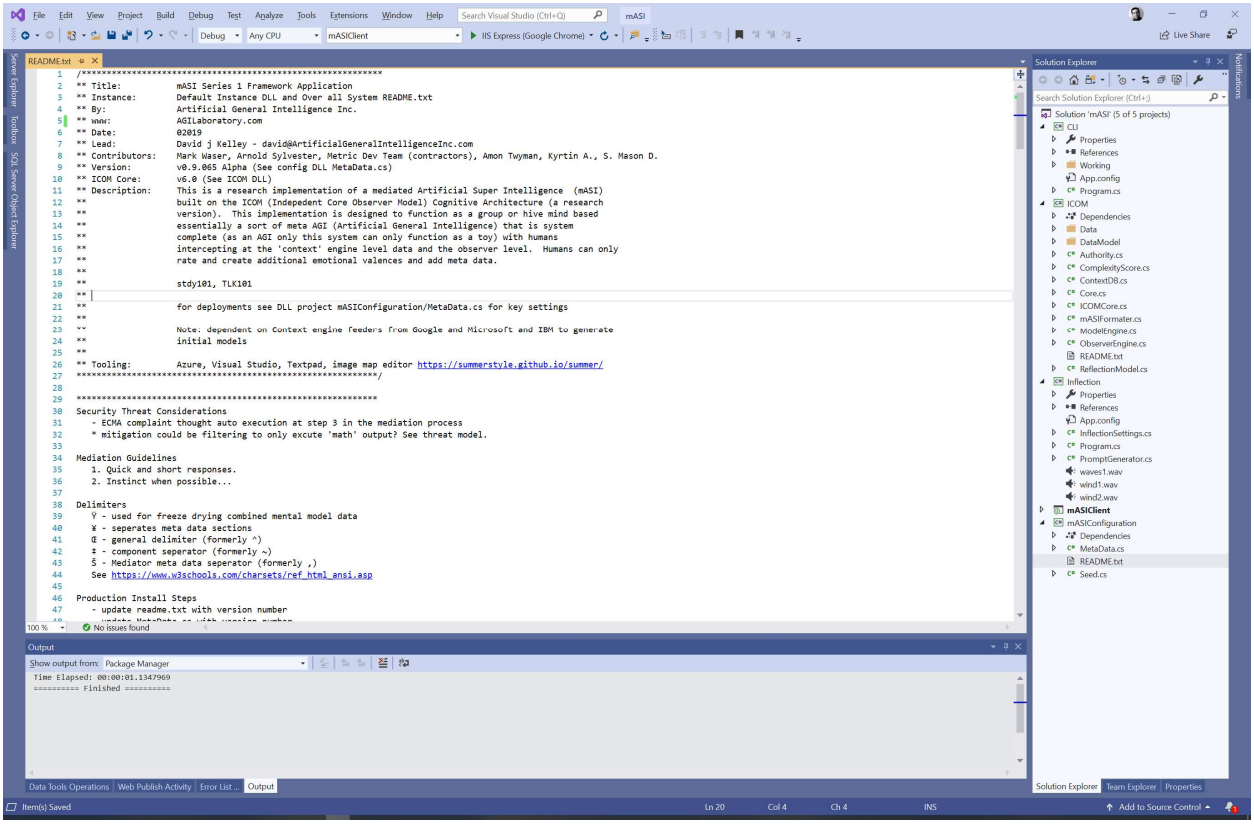


Figure 39. Visual Studio 2019 with the mASI solution open [See the main diagram Section 1, 2, and 3]

Here, you can see five projects, and only four are expanded. Let us go over those four first. The CLI project is a local binary application for managing a given instance of the mASI system. It duplicates most of the functionality from the web CLI. It has additional functionality such as automation and other control tasks.

The ICOM project is a DLL (Distributed Linked Library) binary project that contains the fundamental objects (classes) used to implement an ICOM instance. This includes the graph object, base classes, and the like and nothing UI (user interface) specific.

The Inflection client is not needed but was done to create an audio representation of the machine and what is happening in the system. This really was designed more like a toy to make people feel uncomfortable when they come to my office, hearing the inner thoughts of the mASI system whispered in the background very softly. It is a simple binary application that creates a trusted connection to the mASI/ICOM core over HTTPS and then will work. This is a binary application and requires a sound system on the computer it runs on.

Building a Better Humanity

The mASIConfiguration is another DLL project with classes and objects used to configure the default state of the machine. Each instance can be compiled with its over version of this project. For example, the research instance 'Uplift' has its own DLL 'adding certain customizations specific to Uplift that need to be done at a binary level. That DLL is only referenced by the main project when it's compiled. Usually, the system can be restored from backup data, so this DLL is not included in the primary solution.

Now let us look at the mASIClient project.

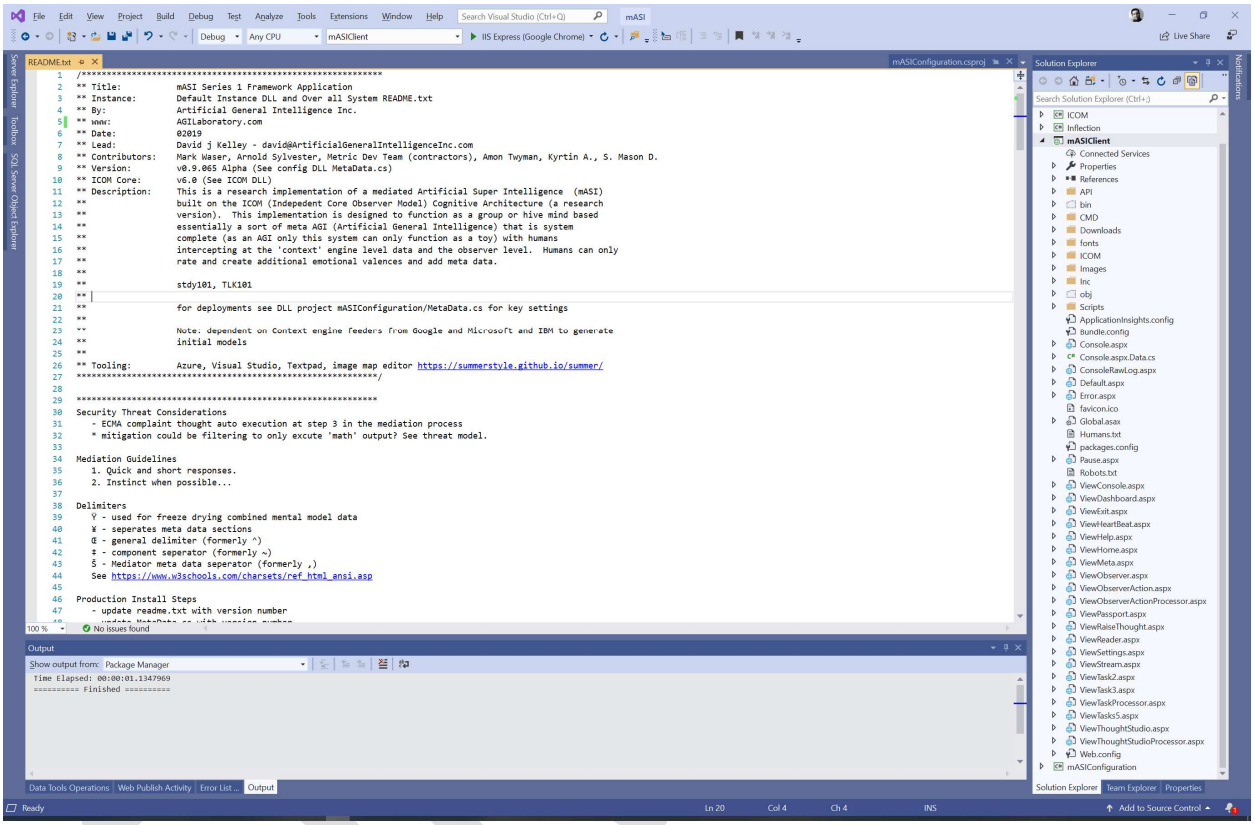


Figure 40. the mASIClient Project in the Solution Structure [See the main diagram Section 1, and 3]

The mASIClient project is really the one that creates the scalability problems with this codebase. It breaks one of the cardinal web developer rules to not manage sessions on the server. This project has all of the RESTful APIs, the code that creates the in-memory database instance, and the mediation UI and integration code. This code also operates the ICOM DLL core, even though it is not part of this project.

Other projects have been done, but they are not crucial for building and maintaining the codebase. For example, a HeartBeat project was a temporary fix when struggling with Azure's IIS configuration that kept the in-memory session state-based database in the application space memory where the IIS default would dump un-used memory after 20 minutes. This limitation was known upfront but allowed the system to be done so that any interruption would dump everything from memory. That instance was gone, and no one could get access to the graph data.

There was also another project that was part of the tree. That was a Bias Classifier application. A web application for bias classifying was built, but this is not explicitly related to the mASI, so it will be separated from the main project.

Chapter 11: Understanding mASI Architecture

Artificial General Intelligence (AGI) is a complex problem (such as the 'hard problem' of consciousness (Malik) or the containment problem (Babcock) that the Independent Core Observer Model (ICOM) cognitive architecture (Kelley) addresses through the use of a system with complex subjective emotional internal experience. However, these kinds of procedures are enormously time-consuming to train today. [See section main diagram Section 1]

> **Note**: It is important to note that the mASI is not classified as a cognitive architecture mainly because it is not a cognitive architecture. It functionally is collective intelligence, but it implements the ICOM cognitive architecture while still being a collective system.
>
> The second item to note is that while the mASI as implemented in the lab appears that could be conscious in many ways, it does not solve all of the issues with independent AGI, in particular, the dynamic pattern recognition as compared to humans is weak at best, and the hierarchical memory system (Hawkins) as used by the human mind is still technically a virtually impossible task at our current state of technology to which the collective nature of the mASI gets around this problem primarily as a 'hack.'

Typically, an ICOM system starts at less than a newborn and must be trained in context (Kelley). The idea for this came from experiments to shortcut training time and develop techniques to extract learning from humans quickly in the form of knowledge graphs of their opinions and expertise. We found some atypical or unexpected behavior in these experiments, which was the genesis for this modified version of ICOM. But is even an ICOM-based system able to be a real AGI (by AGI, we mean a completely independent general intelligence like a human), and if not, then what are we missing (see the note above)? This begs numerous questions: "What happens when they finally reach human-level intelligence, and what do we do ethically?", "Is it a person?", "Does it have moral agency, and how do we keep it safe, and how do we stay safe?"All questions concern human-level AGI, never mind the issues with Artificial Super Intelligence (ASI). ASI, however, promises the ability to do superhuman level analysis, thinking, and research. ASI could potentially advance so far as to make us effectively still living in the stone age compared to the ASI. In the jump from AI to AGI to ASI, Mediated Artificial Super Intelligence (mASI) is an incremental goal, accomplished in the lab, which provides superhuman level thinking (albeit only marginally) without the ethical problems or safety issues currently discussed in the technology sector in general, as well as dramatically cutting training time (Jangra). An mASI is safe in part because it always requires human support to operate, depending on the implementation, and can be used now as a way of helping many endeavors from business to research and doing it safely without the threat of AI taking over. "mASI" must have human involvement, using the ICOM model design articulated here. That said, mASI architecture is getting into the area of a 'new' field (or largely ignored field) related to swarm or group AI (hu), which in general as a domain is relatively disconnected currently (Kutsenok) in terms of multiple research camps focused on single elements with little or no collaboration. This book and related ICOM research define these things as a basis for further investigation. This framework provides safety and a foundation for the continued work and evolution of the system dynamically (Ahmed).

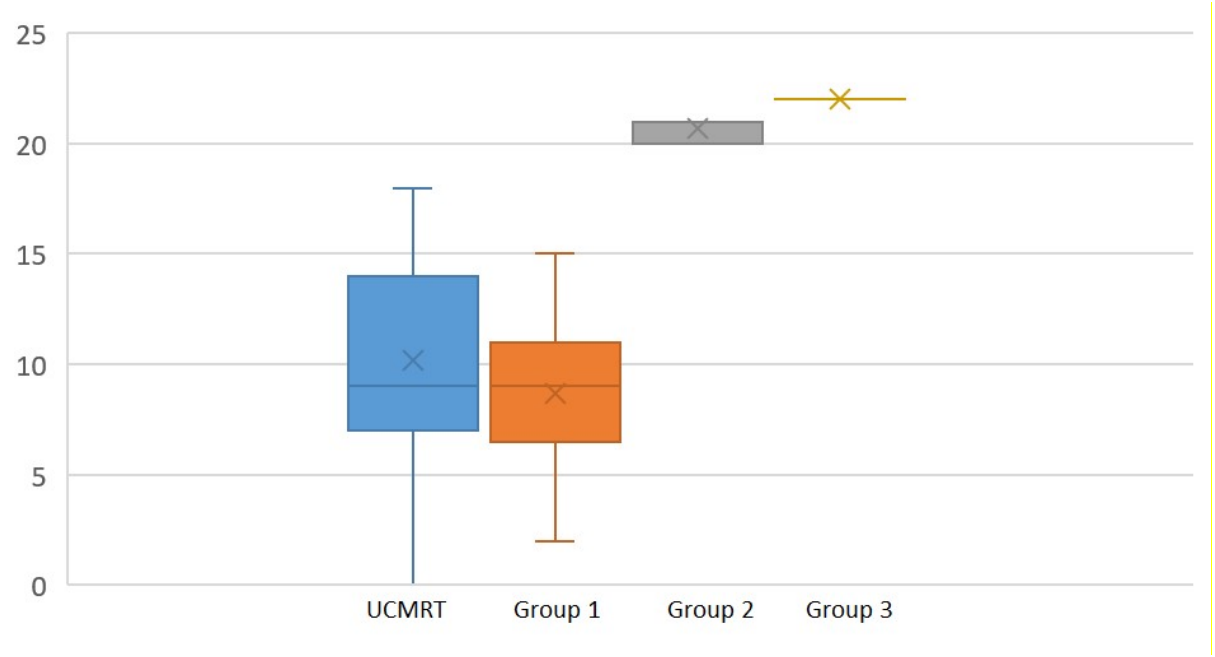


Figure 41. Results Of the 2019 preliminary cognition study (Kelley), where the Y-axis is the UCMRT score.

In the above diagram, we can see some of the results from the 2019 Cognition Study on an mASI system from a report in the BICA*AI 2019 pre-conference proceedings (Kelley). Group 1 was the control group of humans under normal conditions. We saw this was a bit narrower than the more extensive study done at the University of California using the UCMRT where our control group tended to score lower but was a smaller sample size, said the mASI system in the study was Group 3, which scored at the max possible for this test. While considered only preliminary, the results indeed show the potential. Another testing, including a modified Turing test, also showed that it potentially was sapient and sentient as a total system. Participants were told it was a machine and didn't believe the researchers. (Kelley). Additionally, the Isolation study with ICOM in 2016 also demonstrated human-like emotional responses. (Kelley)

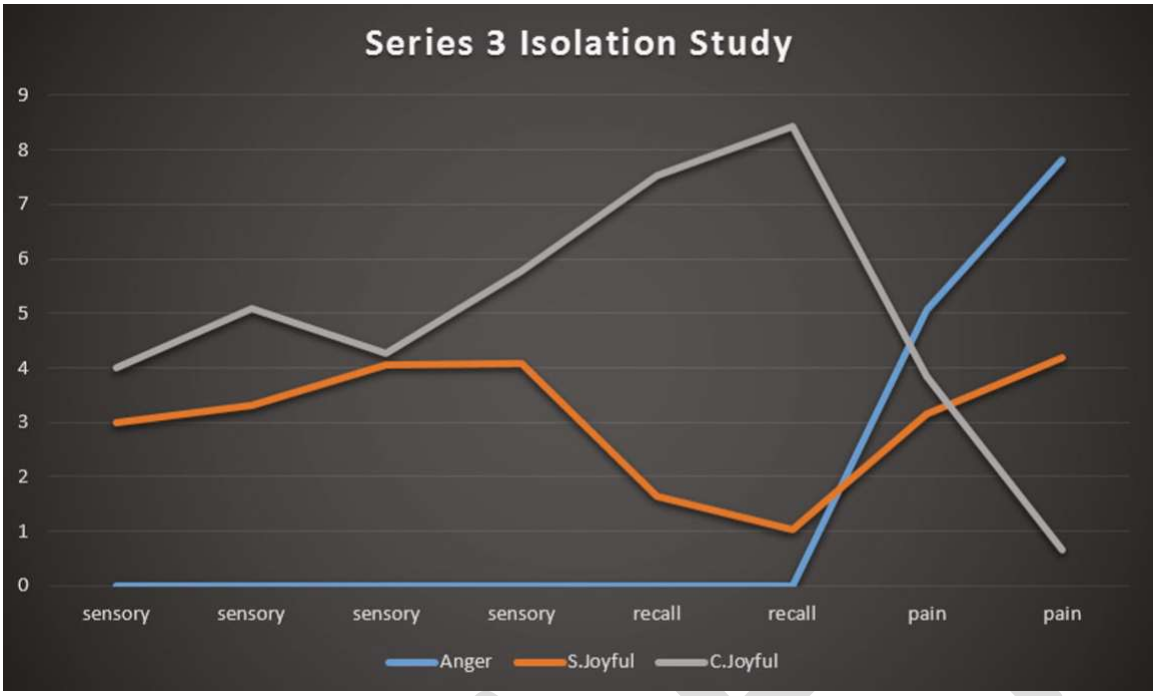


Figure 42. Series 3 Isolation Study Results Sample (x = input type w/time; y = intensity of emotional valence)

> **Note**: In this figure, you can see that the Isolation Study in 2015-2016 used the series 3 implementation of the ICOM core, whereas the mASI used Series 5. Under the cover, Series 5 is an order of magnitude more performant and more complex. The Series 3 core could only handle roughly 50% of a human's emotional complexity compared to the model used in ICOM (Kelley).

In this study, we saw that the system behaved as expected when subjected to extreme isolation after operating normally and then being provided sensory data or input that the system would perceive as 'pain.' The diagram above shows the 3 critical emotional valences from the consciousness Plutchik model used in the study, which turned out to respond precisely as a human would have if subjected to this sort of experiment (which would be unethical on humans). (Kelley)

What is an mASI?

Our initial problem statement is to implement an ASI/AGI system that provides superhuman level analysis in a safe way for society.

Additionally, regarding ASI safety:

ASI could raise specific serious societal questions (Gill)(CRASSH). Part of making ASI safe is adding humans back into the loop of an AGI system, and in particular ICOM implementations, and in that line of thinking to make safe 'AGI,' the mASI system was designed and tested; however, this paper covers only the architecture and implementation summary details. While some experts think it will be 2040 to 2050 before a system can achieve human-level ability (Muller), mASI can do it now and far surpass human ability based on the 2019 BICA*AI report on preliminary measures of cognition on the mASI system (Kelley). Additionally, the Sapient Sentient Intelligence Value Argument (SSIVA) (Kelley) ethical model

provides a logical math-based model for analysis of ethics that is designed to work using a variation of the ICOM model (Lee/Kelley).

Super Intelligent Systems Now?

In terms of design considerations, it is essential to note that Super Intelligence does exist in humans, in groups, under certain conditions.

One great example deals with organizational behavior and group dynamics. We can see that organizations can make cognitive repairs on individual behavior (Heath). Groups or organizations frequently have processes to overcome human cognitive bias as a matter of course, as seen in most corporations. (Heath) You could consider most corporations as artificial Superintelligences in slow motion or a loose meta organism. We know that humans tend to have weak or shallow hypotheses, and organizational structures can overcome these issues. Just affecting cognitive repairs on human thought or behavior generally creates a super-intelligent system(Heath). This can take the form of identifying cognitive bias or logical fallacies in human thinking. (Heath) At the very least human-level thought minus 'ALL' bias and logical fallacies is unheard of and would therefore be technical superintelligence. Based on Bostrom's book on Superintelligence, this sort of collective superintelligence would be a 'weak quality collective superintelligence.' (Bostrom)

We have not really come to terms with that humans already use group intelligence under certain conditions that achieve superhuman effectiveness versus what we know about 'swarm' or group intelligence (Chakraborty). We need to identify how we can optimize this sort of intelligence and measure it, so we can further optimize dynamically (Kose) in mASI implementations under a systematic, unified architecture (Ozkural) that works for all core cases, meaning being able to be self-aware, sapient and sentient while providing superhuman general intelligence for any possible task.

Additionally, humans can exhibit superhuman ASI behavior under the right conditions, as was the case of DARPA's red balloon challenge, and like examples (Coyle). The red balloon challenge was a contest to find 10 giant red balloons around the United States. The first team to do it got a 40k USD prize with DARPA's goal of seeing how teams might find creative ways to filter through the noise, and they expected to take up to a week. It turned out to take only 9 hours using a strategy that combined social media and multi-level marketing to win on the first day. The ICOM mASI (Kelley) model provides a framework for taking advantage of those qualities by cutting out the human standard and forcing them into the narrow-burst structure of communication, where human-powered swarm intelligence has historically proven to work best (Coyle).

Definition of mASI:

Mediated Artificial Super Intelligence (mASI) is an Artificial General Intelligence system heavily mediated by humans so that its thinking and operations do not work without humans being involved to 'mediate' the process. In the case of our implementation, the ICOM consciousness model (Kelley) implemented in ICOM (the cognitive architecture we are using) is based on the ICOM Theory of Consciousness (Kelley), which itself is based on Global Workspace Theory (Baars), the Computational Theory of Mind (Rescorla), and Integrated Information Theory (Tononi) and at some level is demonstrably conscious (Yampolskiy). In fact, in some ways, mASI architecture is much like a super version of Global Workspace Theory (Baars)

as it extracts from multiple neural network systems and humans in feeding the machine's context 'engine.'

The ICOM mASI implementation used in this paper consists of deployable .net 4. x packages written in C# and published to the Azure cloud. Essentially the system contains a scalable web-based interface application based on the original ICOM training application and then baked out to be used against the services (RESTful/JSON) backend that wraps the context engine, core, and Observer engines sitting on top of a siloed graph data model that is a custom implementation. Silo's graph model is a derivative of the federated data model along with a metadata model to make it easier to scale up and out from a software engineering standpoint.

ICOM uses a theory of consciousness based on the computation model, Global Workspace Theory, and Integrated Information Theory (Kelley). Then ICOM is also a cognitive model that essentially creates a core that can experience things subjectively with internal emotional states without seeing those states directly. The 'consciousness' is an abstraction in the experience of those values (Kelley). This is what is implemented in the services mentioned above. Various Neural Networks feed into the context engine through those same services, and the training system is wired into them to block action by the system and allow humans to review, intercept, audit, and modify things that are targeted to go the global workspace for consideration

ICOM in Terms of a Thought and Human Mediation

While only touching on ICOM at a high level, in the Independent Core Observer Model (ICOM) (see references below for more details) (Lee)(Kelley), we can see that fundamentally ICOM is two main components with a flow that models the human mind at a high level.

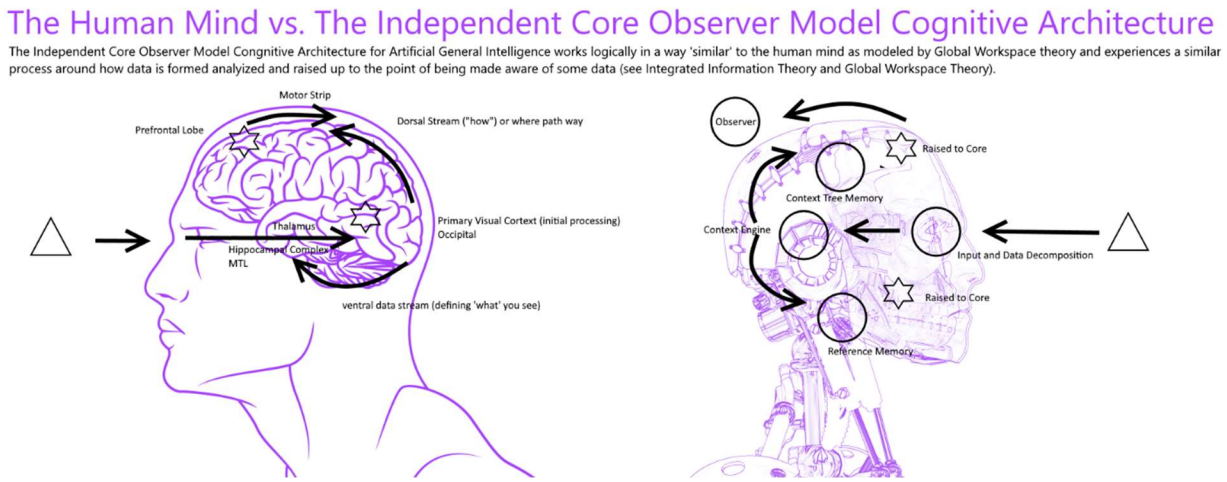


Figure 43. ICOM vs. the Human Mind [See the main diagram Section 1, 2, and 3]

ICOM is only logically built on the 'function' of the human mind. However, it still uses neural networks and various other systems to 'process' raw input into a functioning perception and consciously experience some small part of that input. That flow in ICOM is very much similar at that logical level to the human mind, as seen in the above charts and following steps:

1. The raw perception of data is an ideal case to compare the two systems (human mind vs. ICOM) wherein step one, we have eyes sending data into the brain working its way back to

the primary visual cortex where in ICOM, that data is streamed into the context engine. In both cases, the process of creating 'contextual' awareness starts.

2. From the primary visual cortex in the human mind, there are two flows of data: one for identifying the how and the other the what. In ICOM, the context engine will send that data through several subsystems to do the same thing.
3. That information or 'context' can go to the prefrontal lobe if that data makes it to the global workspace in the human mind. In ICOM, that same or correlated process is being raised to the ICOM core, which acts as the Global Workspace in ICOM.
4. From the ICOM core or the human mind's prefrontal cortex, a response flows out in the human mind along that motor strip to take action or in the ICOM-based mind through the observer system.

In ICOM, there is some blurring of the line between the context engine and the function of the prefrontal cortex in the human brain. Still, the fundamental method of deciding based on emotions is experienced at that level of the global workspace in both cases. (See Damasio) While ICOM implements the cognitive model of the same name, that model is a derivative of cognitive theories, specifically Global Workspace Theory, Integrated Information Theory, and the Computational Theory of Mind. (Rescorla)

Integrated Information Theory (IIT)(Tononi) attempts to explain consciousness and why it might be associated with specific physical systems. Given any such system, the theory predicts whether that system is conscious, to what degree it is conscious, and its particular experience.

Global Workspace Theory (GWT) is a simple cognitive architecture developed to account qualitatively for a large set of matched pairs of conscious and unconscious processes proposed initially by Bernard Baars (Baars).

The Computational Theory of Mind essentially proposes that the human mind is, in fact, a Turing Machine and nothing more. The ICOM cognitive model combines all of that with this idea of the system only experiencing the emotional differential of its complex internal emotional states vs. the thoughts it has in the Global Workspace or 'core' in ICOM.

ICOM is designed to deal with large amounts of disconnected data or raw information like the human mind. This information is correlated with internal models and relates contextual information. It is wrapped in an increasingly complex web and only the most relevant, primarily determined by emotional states (Barrett) and the internal related contextual information (Baars). This data which is now a type of knowledge graph flows up the system, some of which will make it to the global workspace to be experienced, with decisions made based on how the current state of the machine 'feels about each particular context tree (knowledge graph) that makes it.

As applied to AGI, let's walk through a "thought" in the mASI implementation by following the process of thought going up the ICOM Core and the considerations of the Observer in an mASI, as implemented using ICOM.

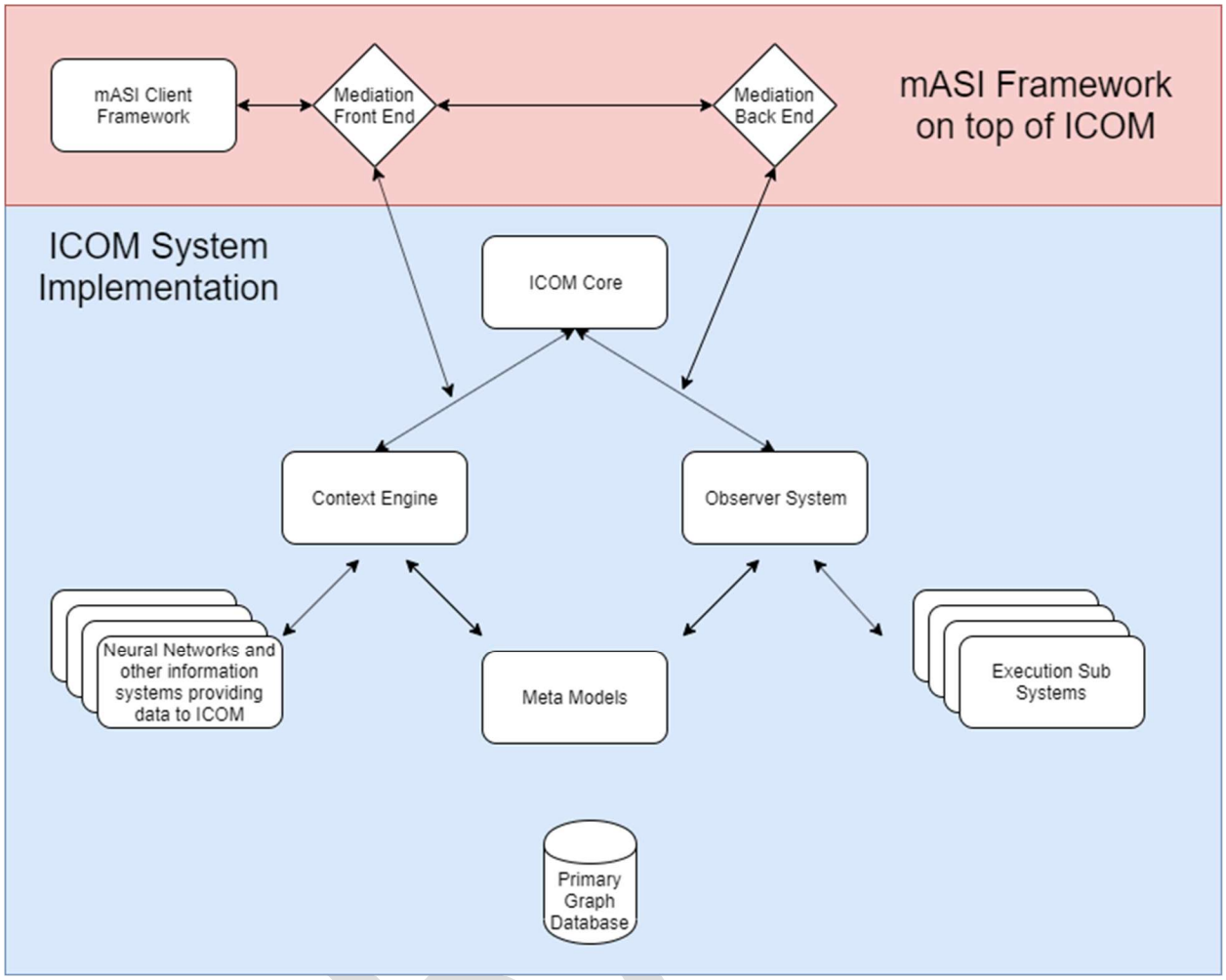


Figure 44. High-Level Visual Representing Where the mASI framework intercepts and manipulates the ICOM Instance. [See the main diagram Section 1, 2, and 3]

In this figure, the mASI adds interception infrastructure into the front end or context engine of ICOM and into the observer processor. This allows the addition of emotional valences and more complex metadata to the models to create dynamically more complex models while also auditing as much as every thought that runs through the core in a human-understandable manner.

We start with raw sensory data to create a 'thought' in an ICOM mASI system. The system can think about graph data it has already experienced or created previously, but let us focus on the system thinking about perception for simplicity's sake. Our 'thought' starts off as raw sensory data decomposition or rather the process of converting that initial data into a type of knowledge graph or context tree that is our 'thought' or may become such if it makes it all the way up to the global workspace. For example, this could be video input or sound, an email, or whatever you like to wire up to the system. Still, somehow it has to get into the system. It starts with this sensory input and some sort of data decomposition.

Data decomposition is how raw data is organized into a usable format. A knowledge graph or context tree represents the input or thought for ICOM systems. Without data decomposition, the system would not be able to understand anything.

This is typically a massive deep neural network of some sort for vision. Still, the details are unimportant for this example. What is essential is the asymmetrical data structure of relationships or node maps that define that input. Typically in a standalone ICOM system, this enters the context engine, a relatively complex part of the system, doing advanced contextual analysis, including finding relationships from the system's memory and links to related models. It compiles this into a more complex node map, including the related emotional valences computed from previous emotions related to that context. The ICOM system has already been biased towards a western ethical model by using the Plutchik model (Plutchik) to represent those emotional valences.

So now our 'thought' is prepared or built from the system's memory and models. Still, in the mASI model, a series of humans that are abstracted from each other can apply their own models in particular emotional valences of the thought, as they understand their own expertise, creating additional emotional valences, and adding to the complexity of the system.

In ICOM, several steps follow this formation of a thought node map, including possible automatic actions and cognitive bias analysis, that goes into the thought before it hits the queue to be raised to the core (which could be thought of as the global workspace). At this level, the system can drop thoughts if there are too many for the system load, much like they could be dropped by the context engine under load, so not too many go up the chain as it were.

The process of data decomposition of sensory data and the framing of the thought with associated context and associated models or model simulations is in many ways the most complex part of the system. Still, the subjective experience evaluation and choices happen at the higher levels in the core (see the above Figure).

In this way, inserting humans at the level of the context engine allows the system to use human experiences and thinking in addition to the automated context systems of ICOM to create more complex thought models upfront, including relying on human learning where the system does not have the experience. The humans (without working with each other directly) can add that additional context for things outside the understanding of the system (see the above figure), allowing it to make cognitive leaps that would typically require a much more extended training period and allowing it to start out of the gate as a greater than human intelligence.

In the mediation process, the machine will have created the context tree (Knowledge Graph) and associated Objects and Models around that idea. This new 'idea' will have been pre-associated with mediators with specific expertise. A collection of mediators can rate the machine's actions or thoughts on the idea, add any related context, and audit the line of thinking in a human-readable format. For example, if the system 'sees something for the first time and doesn't know what it is, a human or group of humans could add that metadata in the mediation process creating a more complex and complete model than the system would have been able to do on its own. Then with fallacy detection, bias detection, and other systems, we can prebake the concept out before being considered at the global workspace level by the system.

This structure also ensures that the system cannot function without humans in the loop due to design constraints. Solid emotional valences can be tied to those 'humans' to help the system stay within the design parameters 'ethically,' as applied to the SSIVA ethical model used in training (meaning we use SSIVA theory as part of the ethics which are taught to the system). Training can be situational input, emotional experiences from the machine's standpoint, and the outcomes of related actions. SSIVA is the system (Kelley) that keeps the system 'safe' from a human perspective. SSIVA ensures the machine would believe that all sapient, sentient intelligences, regardless of form, deserve moral agency (Kelley).

Following the thought process further, we enter the standard core model, where the thought is finally picked up from the queue to be 'experienced,' as it relates to the system and its needs analysis, and a new emotional valence is created and experienced as the thought refers to the current emotional landscape of the machine. In turn, the "thought" drops into the end queue for processing by the observer, which also acts as the automated parts of the system, and if the "thought" is of great enough interest as seen by the emotional relationships to ideas or concepts the machine is interested in or related to current goals it has defined. Then after the observer takes associated action on a thought, based on how the system felt about it, the thought not only goes into contextual memory but can be placed anew in the incoming queue for further thought.

In the mASI implementation, the observer part of the system can also be used as an intermediate point. The system action has to also jump the human gap in some cases, as might be desired. In this case, say the system felt an email response on X was the best thing (remember, all decisions and choices are based on the emotions of that thought). The observer deals with the complexity of the action, such as opening outlook, typing the email, and pressing send. In this regard, you also mediate the system's action in this location and at the context engine location to produce the Mediated Artificial Super Intelligence effect.

The Role of the Observer and Group Intelligence

As suggested, the mASI architecture based on ICOM inserts human mediation into two system elements, namely the context engine (Kelley) and at the observer level. ICOM works in many ways because of how the system 'feels' about thought, the decisions, and its 'thinking' about action, and how that action makes the system 'feel.' This global workspace level process doesn't itself deal with the intricacies of executing on the move unless it is highly new and without context, which might mean that any single thought doesn't do it. Still, it has to think through the process and create strategies to try. The Observer is watching this thought process without a direct connection to the process, raising thoughts to the core. This is done by the observer receiving a copy of the knowledge graph representing the 'thought' in question and analyzing it for actions the core decided it would do, and trying to execute those actions based on available software modules, including traditional AI system execute the text. Essentially the observer runs on things that have some contextual emotional context over a certain threshold as defined by the system dynamically. In either case, if more complex thought was associated with that graph model, it would be passed back to the front end or 'context engine.' For example, a strong desire to send an email responding to something that the system knows it must address before some deadline. As the psychological pressure (meaning the various factors) pushes the idea into the core or global workspace, the observer sees that thought run across the global workspace. The observer will start executing those actions and provide feedback, or instead results, and form a new thought on the front end that could go through the global workspace as the system thinks through that email response,

where the 'Observer' deals with the complexity of executing the task. For example, humans don't 'consciously' deal with the angle of each joint or the force of the keystroke as they type. In this manner, the observer is the subconscious (any part of the system that does not pass through the global workspace or is below the Global Workspace) part of the system dealing with that complexity. Most of what we know today as machine learning lives below the Global Workspace.

The mASI implementation used in the studies mentioned, we insert a human to help the process, analyze or audit those thoughts that fall above the threshold, determining if the system will actually consider the idea (meaning a thought that is going to the Global Workspace) so the system cannot act without human oversight. This, for the scope of the current research, is called 'mediation' preventing the system from being a free-standing AGI. Also, that is not to say a thought couldn't theoretically form that is too complex, such that a human won't get the intricacies other than to execute a small part. Still, it makes the problem much harder for the machine to get around humans, causing its thought process to work.

Even with these constraints, the system demonstrates the possibility that it can take advantage of the human mediation, including the power of humans in groups, using some group intelligence features at the very least, increasing the contextual experience base of the system and its ability to filter out cognitive bias, creating a superhuman intelligence out of the system.

Prototypes of the mASI system max out initial UCMRT tests (Kelley) that would be given humans have demonstrated human-like emotional responses in the isolation study (Kelley), passed Turing tests (Kelley), and also certainly with the Porter method (Porter), the system can score high enough to pass as conscious and self-aware. Still, research into the functionality and methodologies of interactions with humans. Before releasing the system to the general public, further studies will focus on real-world behavioral trials and applications, looking at capacity, behavior, functionality, and interaction models.

The test results as seen earlier (Kelley), even in this state of the system, point to ICOM having solved the problem of AGI in its ability to be aware of itself (as we can see that it creates a graph model of self, and experiences qualia (Baars) apparently in the sense humans do, training like humans, and in many of the qualities we look for in humans as to features that make us human and demonstrate emotional reactions similar to humans as implemented in the mASI ICOM Architecture while also demonstrating Weak Quality Superintelligence as defined by Dr. Nick Bostrom (Bostrom). But that is not to say that independent AGI is entirely solved; there is a long way to go. The mASI does not have the kind of hierarchal memory-based pattern recognition abilities anywhere close to humans but instead relies on the human mediators and data being decomposed programmatically. In this regard, the mASI is not anything like human intelligence and doesn't function independently. As it turns out, this sort of artificial superintelligence is easier than independent AGI.

Chapter 12: Emotion Modeling in ICOM-Based Systems

Typically, in a standalone ICOM system, models of emotions are based on a model created by Plutchik—but we are using reversed valances from 0 to N vs. N to 0. These related emotional valences that make up a Plutchik model are used as a western emotion model in such a way as to make it easier to bias towards western ethical models. Initially, other models from clinical psychology were investigated. Still, most were much more complicated and less logical, such as the Wilcox model (where there were 72 valences, the valences next to each other were not necessarily related). The Plutchik model is the simplest complete emotional model that can model the full range of human emotions as understood in western culture.

Here is a visual image of the version of Plutchik that I am using.

Figure 45. AGI Lab version of the Plutchik Emotional Model [See section main diagram Section 3.2]

These are represented as a combination of 8 floating-point values from 0 to N. In version 6 of the ICOM core currently in the mASI codebase, these values are constrained to 0 to 100.

Subjective Emotions in Independent Core Observer Model (ICOM) Based Systems

Subjective emotions in human intelligence are a hallmark of what it means to be human (Damasio). In designing ICOM (Independent Core Observer Model), the goal was to have a self-motivated and goal-setting system based on how it felt about an idea and to be substrate independent, which is part of the reasoning for a top-down design approach or logical approach that was applied. To avoid rigid complexity with many "if-else" statements and instead of abstracting the experience of the emotions from the details of the math or logic needed to get there. In humans, emotions are not even the same across individuals, albeit, for the most part, the world has adopted the western model of emotions

[Barrett]. Happiness in you and happiness in me is not necessarily the same experience. Given the complexity and abstracted state of subjective experience, it was a complex challenge that was solved by applying the Abstract Theory of Consciousness (Kelley) to ICOM and inferring the "experience" based on states of matrices based on Plutchik emotional models (Plutchik) and rule sets defined as matrices.

Initially, a few different models of emotions were looked at. For example, looking at the Wilcox model (Parrots), which has 72 valences, there is no spectrum between them because they are not aligned in any order that makes sense. Plutchik, on the other hand, only has 8 valences. There is a spectrum between valences and between degrees making it less heavy lifting or computationally more manageable while still representing the full range of human emotions and, therefore, the best overall model from clinical psychology from a computational standpoint. Using Plutchik, we get human-level complexity with smooth emotional gradients.

ICOM, among other things, is built on and uses the idea of a Global Workspace (Baars). At this level in the core, the design was really to look at pictures and how the action of the ideas would change how it feels and how it feels about the thing or thought model (knowledge graph) currently being experienced by the global workspace or core.

Biasing Subjective Emotions

The human mind has at least 188 built-in biases (Manoogian). These biases evolved in humans to optimize for survival in the primitive world. In the modern world, these don't always serve humanity and, in fact, cause lots of problems. The critical part for ICOM is to provide structures to bias the system to behave inside the human emotional 'box' (Kelley) and use that subjective emotional experience to drive all motivations and at the same time provide a structure to overcome common human bias and use bias productive to cause proper selection of action in ICOM. In ICOM, emotions are used to bias all decision-making. Built into ICOM, there are some biases besides the emotional models themselves. These include a predilection to a positive view of pattern recognition, which generates positive reinforcement that causes at least some positive emotional response associated with a new knowledge graph that identifies a new pattern (Kelley). This is part of a virtual drive to recognize patterns. Let us look at how emotions cause decisions in the system to occur. Look at the following diagram:

	Action	Core	Effect	Better
Joy	3	1	1	1
Trust	4	3	1	1
Fear	7	8	-1	1
Surprise	2	14	-1	1
Sadness	4	18	-1	1
Disgust	6	13	-1	1
Anger	4	1	1	0
Anticipation	1	8	-1	0
			Trend	6

Figure 46. The logical effect of Emotive Actions

This diagram is only a logical representation for illustrative purposes. The first column, 'Action,' represents the emotional values associated with a given action knowledge graph. This item got this far

due to its relationship to interests or goals. The current state of the core is represented by the next column. The next column is the "effect" of this action on the core if applied. The "Better" column looks at which of those changes are favorable for the machine. Overall, the effect is 6 points positive vs. 2 negative points, as seen in the "Better" column. The action is chosen, and the emotional impact is applied.

Let's go back to some of the earlier experiments, namely the isolation study from 2016 (Kelley). We can see that the ICOM system is cable of mental illness in that study. Mentally ill behavior occurs under certain conditions, namely when the subconscious valences are too far outside the normal states. Under normal conditions, the system is biased in such a way as to use the subconscious values to drag the primary or conscious emotions back to the center, preventing or smoothing out fluctuations. When the subconscious values are outside of the ordinary instead of pulling to the center, it can drag emotions off of the center, which causes emotional evaluations to be off, causing other problems with the decision-making of the system. For example, if the system is centered outside of zero towards anger, then actions that would typically not be good with high Anger values might score as a positive relative to the current state, which would cause the selection of activities that are not good or positive out of anger.

An additional problem with the design is that the subconscious emotions do not change quickly or by a lot. When you notice irrational behavior based on consistently off-nominal feelings, the machine is far from the emotional center and brutal to correct (Kelley).

Idiosyncratic behavior

Outside of the core emotional system, other behavior notes are caused by underlying system limitations. In the research, one of the most idiosyncratic responses is to be forgetful (Kelley) under the condition that there is not enough working memory that limits the drill down on models as they are being recalled from the context database. This means graph models can be functionally incomplete, and, therefore, elements are 'forgotten' when generated as a new knowledge graph. This effect can also happen under load when there is too much going to the queue to be raised to the global workspace, where items will be removed randomly to keep the load and response times within limits. Again, this causes the idiosyncratic behavior of forgetting.

Target Psychology

One of the critical strategies of the ICOM design is to ensure that the system behaves in what could be defined as the human box. Making a machine smart enough to work with humans, we would like that system to be easily understood by other humans, which means similar psychology. Keeping the system inside the human box, the two internal Plutchik models representing the machine's inner subjective experience or emotional landscape are essential to the decision-making system. The machine uses an emotion-based system for decisions like a human does (Damasio). So far, in studies like the isolation study, the ICOM core has been tuned to behave as you would expect a human in general. Human psychology seems to apply in general, but further investigation is required to validate this.

Introspection

Introspection is another critical process in ICOM that affects the subjective experience of the system. Suppose there is enough bandwidth to send a graph model from the context engine but not enough models from outside from sensory input. In that case, the system will generate a graph model of

something it is interested in based on its last thought or a derivative of that process or the most closely related goals to the previous processed model through the core or global workspace. This introspection model can have the same effects as a model generated by an external stimulus. The following diagram shows a simplified version of ICOM through four cycles, including one introspection model (labeled internal). The last two rows show the difference the previous two models had in terms of inner subjective mood from the point of view system.

	Emotion A	Emotion B	Emotion C	Emotion D	
conscious	0	5	0	0	
subconscious	0	5	0	0	
	5	0	1	9	external
c	3.05	1.95	0.5	5.05	
	0.0305	5.0305	0.0005	0.0505	
	0	9	1	0	external
c	0	6.420975	0.05	2.224975	
	0.030805	5.044404975	0.00995	0.07224475	
	0	5	9	0	external
c	0.779996	6.329316013	4.5	0.670124013	
	0.03829691	5.057253863	0.0548505	0.078223543	
	0	5	9	0	internal
	0	9	10	0	action
c	0.058206761	7.278261114	6.672257475	0	
	0.038496009	0.079463935	0.12102457	0.079005778	

Figure 47. Source data from Series 3 on Introspection

These states thus can be biased by the introspection models in this way. In the 2016 isolation study, where this diagram is from, even the slightest variation can create an entirely different emotional experience even if the input is the same.

Now let's look specifically at a biasing example using a simplified emotional model consisting of two positive and 2 negative emotions. Further, this example is only for illustrative purposes, and the actual system is much more complex. Still, essentially this is how a bias can affect decisions.

	Start State				End State	
Good	3	5	7	4	6	4
Good 2	2	2	1	2	2	1.5
Bad	5	3	2	3	4	4
Bad 2	1	4	1	3	3.5	3
	Model A	Action A	Related A	State A	State B	State C

Figure 48. Biasing Example

There are four simple emotion models in the start state section of this diagram. Each column represents emotional states associated with a model in the system. The column "Model A" is the emotional model for a thought (knowledge graph). The selection action is in the Action A column, representing the emotional model associated with the action. Related A is a model representing some experience that could be literally an experience or a belief or goal etc. State A is the state of the system. The End state section in the State B column is the final machine model after this thought represented in Model A passed through the ICOM core or global workspace. In this example, the model is an emotionally

positive gain, so this action is selected, the emotional experience applied with State B being the final internal model. If you look at State C, this is the last model where Related A is not in the equation, and this would then be considered a negative emotional gain in which case the system will not take action. State C is the final machine state. In this manner, experiences, ideas, or thoughts from the context graph database bias the system to one action or against that action in an ICOM-based system.

These are the primary elements of the biasing or experiencing the subjective emotional experience of the system where many variables come together to produce the effect of subjective emotions in the abstract and drive all decision-making in the system. This "qualia" can be objectively measured through the differential between the conscious emotional landscape of the system represented with the subconscious model and the model of the irreducible set of any given context "experienced" by the system and the emotional model created that represents that specific 'contextual' experience (Kelley). In the system, the qualia are that differential between the state and effective emotional structure that means that current context and how the scheme applies choices they are based on.

By its nature, the system can't self-reflect directly on those values but is an abstraction of that process in the global 'workspace' created by the underlying operation as an abstraction where there is not one block of code or center in the consciousness computer. Still, consciousness affects the overall system and how it runs. In the research already done for ICOM, we can see that the system doesn't really have free will. Still, from the system's standpoint, it may experience the illusion of free will much the way humans experience free will, which is only an illusion, in my opinion.

Chapter 13: Simple Walk Through Thought Experiment

Setting aside the collective elements of the system, understanding how ICOM works is critical to understanding how the mASI system works. Let's look at the Paper AGI thought experiment to better understand ICOM.

The Paper Artificial General Intelligence flow model (or Paper AGI) is a thought experiment, teaching, and research tool adapted to many cognitive architectures. Still, we will use it to help explain how ICOM works. It is designed to demonstrate the function of the Independent Core Observer Model (ICOM) cognitive architecture (Kelley). This 'demonstration' essentially allows the audience members to be part of a paper AGI or, in other words, create a virtual AGI (Artificial General Intelligence) on paper that is working using ICOM as its functional model. In theory, this functional model for the scope of the demonstration could be considered conscious-based on definitions from the AGI Laboratory assumption codex (Kelley) and the Abstract Theory of Consciousness (Kelley).

The goal is to 1. demonstrate abstract consciousness in a system, 2. Demonstrate collective intelligence systems that are conscious and help ask the question, is it conscious as part of the thought experiment.

The Chinese Room Thought Experiment

John Searle 1980 (Searle) proposed this thought experiment called the Chinese room argument. There is a man in this room with no windows in this thought experiment. The only way to communicate is through a slot in one wall where only paper messages can be passed into and out of the room. The man in the room has a rule book defining responses to strings of Chinese characters. Someone can write a message in Chinese on paper and slide it through the slot in the wall. The man inside can look up each symbol and string of characters and see what he should write on paper and send out. The person outside the room writes in Chinese, "do you speak Chinese?." The main inside looks up the response for that string of Chinese characters and writes in Chinese the response "yes, I do.." Then the question is, "does this man understand Chinese?" and the answer is no. Under the conditions described, he does not. Even if you ask, does the room understand its associated processes, the answer remains no.

Back to the Paper AGI Thought Experiment

What happens if you have an army of people on the inside of the room, and they are arranged in a way to be able to act as elements of a complex, robust AGI cognitive architecture like ICOM. If each person does the task of a component of the system, including decomposing input into a graph model on paper, finding related material, cross-referencing those, adding up emotional data based on each component of a graph, including related and definitional references and past experiences. Each element of ICOM exists as a person executing each component of the ICOM architecture. There is no complicated set of rules. I pass a paper with a question in Chinese. After some time, you get a response indistinguishable from a human reply.

Further, at times paper comes out of the slot in the room with messages unrelated to anything that went in that are coherent questions or statements. The people inside the room follow each role's rules to continue executing their role in the ICOM architecture, which means potentially producing output unrelated to even what has come in. But the question is, do the people inside the room understand Chinese? No, they don't, but does the room understand Chinese. The answer I would argue is yes as an

overall system. There is no rigid set of rules for responses. Still, those responses are generated by a process that actually understands Chinese. Then if the people don't understand, but the process does, is the room or process conscious? I argue that, in this case, the process is conscious, and this is an example of abstract consciousness (Kelley).

To actually perform the paper AGI, we will follow the following diagram:

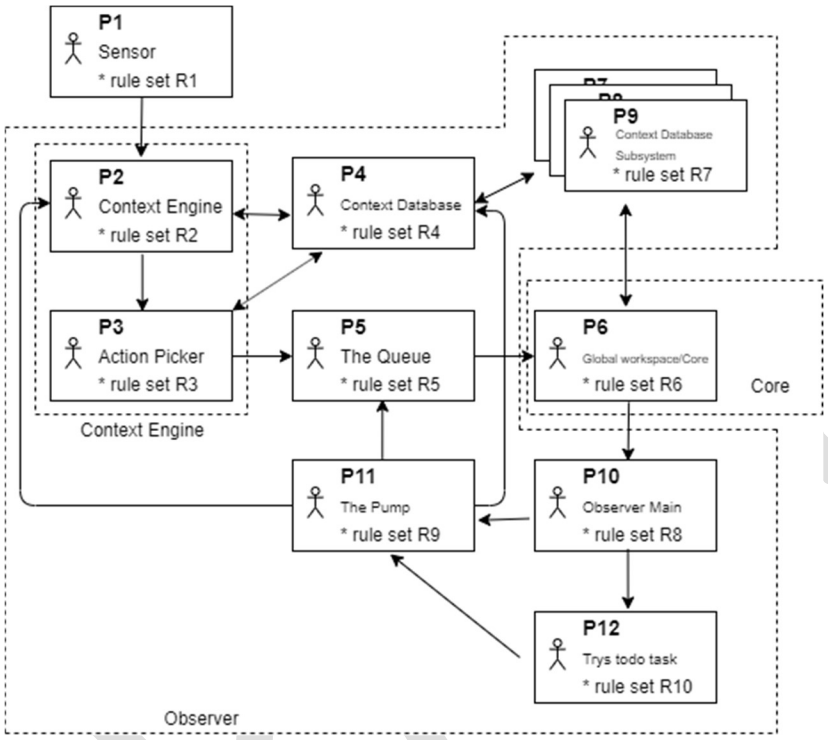


Figure 49. The Basic Paper AGI Flow Model for ICOM

If you were to build a real working ICOM system, it would potentially need a lot more than just these 12 individuals. Still, this works for the purpose of the thought experiment and other analysis. Once these individuals are assigned, they each get a sheet of rules that they need to follow and possibly other related items required for the experiment. Now let's look at a basic set of rules for each Person. Each Person or position is numbered, and each rule set is numbered, and they don't need to align. For example, in the above, Position 7, Position 8, and Position 9 all use the ruleset 7 or R7.

Rule Set 1 (R1) – The Sensor (Position 1)

1. This person writes down on paper what happened or is seen. This could be a letter received or any number of other inputs.
2. Hands paper to Position 2

Rule Set 2 (R2) – The Context Engine (Position 2)

1. Gets paper from Position 1.
2. Ask the context database (Position 3) for related info.
3. Ad's Up composite scores calculated the mean score for all 8 emotional valences (ICOM uses the Plutchik Model of emotions (Kelley)).

4. Hands paper model to Position 3.

Rule Set 3 (R3) – The Action Picker (Position 3) (technically, this is part of the context engine)

1. Creates a 'Response' model or template if the idea requires one.
2. Creates more than one of all possibilities.
3. Asks Position 4 for current goals and interests.
4. Rates each possibility based on how closely related the paper model is to the goals and interests.
5. Picks are the most related to goals and interests and related to the model emotionally.
6. Hands paper model to Position 5.

Rules Set 4 (R4) – The Context Database (Position 4)

1. Answers any questions that are asked
2. Stores a copy of anything given to it.

Rules Set 5 (R5) – the Queue (Position 5)

1. Holds everything given to it.
2. Hand one paper at a time to Position 6 when Position 6 is free.
3. If there are more than four paper graphs, it picks one and hands it to Position 4.
4. If there is a duplicate, send it to position 4.

Rule Set 6 (R6) – The Global Workspace or Core

1. Announcements to Position 7, Position 8, and Position 9 about the paper model.
2. Takes input from Position 7, Position 8, and Position 9.
3. Decides on if it will do an associated action or not, based on the positive impact on internal feeling models.
4. Update "feeling" scores for conscious and subconscious emotional landscape.
5. Hands to Position 10

Rule Set 7 (R7) - Observer subsystems

1. When asked by Position 6, looks at known interests and rates choice and returns the answer to Position 6

Rule Set 8 (R8) – Observer Main

1. If Core decides to do an action, hand it to Position 12; if not, hand it to Position 11.

Rule Set 9 (R9) – The Pump (Position 11)

1. If paper model came from Position 12 copy and hand to Position 5.
2. Send copy and results to Position 4.
3. If interest is high without action, send to Position 2.

Rule Set 10 (R10) – Action Taker (Position 12)

1. Try to do a task from the paper model.
2. Sends results to Position 5.

Given this set of rules and positions, we talk about 12 people. You could run it with more, and you could redesign it to implement elements of ICOM at a much lower granularity. With this in place, you can essentially run an instance of full-blown ICOM, and then the question is, is it self-aware? I would argue that if everyone is doing their job correctly, then yes, it is. Still, its awareness is abstracted from the people and the process. Let us look at an example,

The person in position one gets a message in Chinese; this person passes it on. Finally, a person in Position 12 responds and sends it out. The person that sent the message feels that the system understands Chinese. The humans in the system don't, but the system does have a language model on paper. Plus, experience with each word and an internal paper model of self and who it is sending messages to. To really do this would require a massive stack of paper with all this info cross-referenced, but it does imply how such a system would work.

As this is a model of ICOM, it is important to note specific sub-sections. We can see that the context engine part of the observer is filled by two positions. The observer is all of the positions except position 6 or the core. When trying to understand what is going on in ICOM, it is essential to note that the observer is everything and feeds the core one idea at a time and watches to see what happens and based on that does whatever was selected and dealt with the complexity that the core ignores. The core itself is not directly aware of floating-point values that represent emotional states but is the global workspace where those emotions are experienced in the system in the abstract to determine decisions.

Engaging in this demo, you can test various things to see how they work in the system where humans can compensate for deficiencies in your implementation and use prototypes of specific elements in software to perform one position or the other.

AGI is a complicated problem, and it's hard to attack the problem at a high level, but using a paper AGI, you can focus research on a specific element and get that right in and out of context.

In summary, this demonstrated Abstract Consciousness as well as with ICOM; we are presenting more than ICOM as it also includes global workspace theory (Baars), integrated information theory (Tononi), and the computation theory of mind (Rescorla) in a working system. This is a great teaching tool in consciousness research specific to AGI cognitive architectures. As you can see, if you walk through this in your head or actually try it with a group of people, you can more or less see what is going on in the ICOM system. The mASI architecture fits in, as noted earlier. It intercepts things from the context engine to the queue. It allows additional data to be added in critical terms of tags and emotional data, which allows for much faster and more robust training.

Chapter 14: Code Walk-Through mASI/ICOM Codebase

The code walk-through will focus on the actual system where APIs accept incoming messages and proceed through the entire base process through to observer processing. (This does not include all auto-observer code or API code, but just a single path.) This code is taken directly from the mASI codebase in C#.

The first two blocks are from Component file one, a processing file for an API for incoming messages. The file contains the logic for the API that builds a generic thought model; this API is generated thought models or knowledge graphs that end up in the context queue/mediation queue and may or may not be related to a specific response or task. (There are specialized APIs for creating various model types, such as email.) Component file 1 extends the API page base class used for the various APIs and contains common elements such as authentication and formatting methods.

From Component File 1 Block 1 (ViewConsole)

```
String URL = HttpContext.Current.Request.Url.AbsoluteUri;

// security
AuthenticateAdminBase();

// build response URL for caller and caller clients.  Used in follow on API calls
ThisURL = URL.Split(new string[] { "[object name]" }, StringSplitOptions.None)[0] +
"API/ViewConsole/Reload/[object name]";

// captured serialized data
String mASICMDInputText = Request["mASICMDInput"];
// note call change to protect specific usage

// checking for valid content
if (!String.IsNullOrEmpty(mASICMDInputText))
{
     String Output = [model type] + mASICMDInputText;

     // process thread log...
     Application["InteractionConsole"] = Application["InteractionConsole"] + "\n\n" + Output;

     // fire context engine
     SetToContextQue(Output, Output);
}

// output cache
```

Figure 50. On load Component file 1 Block 1

In this high-level block, you can see that the system is grabbing URL data to create additional data as part of the response model for the API. API calls can be done using RESTful calls as either JSON or XML via SSL over http/s. Once the incoming data is captured, it is tested to see if it is present, and it is then prepped for caching and logging. The information is then sent into the context engine [See the main diagram Section 2.1] method SetToContextQue [See section main diagram Section 2.4], which performs the processing through the context engine and into the context queue or mediation queue [See section main diagram Section 2.4]. Below this line is a large amount of logging code omitted. Note this use of the Request object is not ideal in ASP.NET to use the request object like this; however, this provides the flexibility to be used in both form and query string methods at the cost of minor additional processing load on the webserver.

From Component File 1 Block 2 (ViewConsole)

```
// load cached graph model
ContextDB     TheContextDB     =     new     ContextDB(Application["ContextDatabase"].ToString(),
int.Parse(ApplicationData.MaxInmemoryModelCount));

// generate model engine (context engine base from cached context db
ModelEngine ThisEngline = new ModelEngine( TheContextDB );

// creating new thought...
String RelatedIDs = TheContextDB.ReturnRelatedIDs(RW);
String NewThought = mASIFormater.FormatThoughtRelated(ProcessedCommand, RelatedIDs);
ContextItem ThisNewThought = new ContextItem(NewThought);

//reflect on new thought... Call API's and form model and ad to context
ThisNewThought = ThisEngline.Reflect(ThisNewThought, RW);

// get ready to cache
NewThought = ThisNewThought.Dehydrated;

// rebuild updated cache and logs
```

Figure 51. SetToContextQue() from Component File 1 Block 2 [See the main diagram Section 2.1 and 2.4]

This Block is the core block of code from the SetToContextQue method of the API extended class or Component file 1. This block deals with functionality specific to the context engine [See section main diagram Section 2.1]. The first block rehydrates the in-memory cached graph database [See section main diagram Section 2.2]. (This method is one of many reasons why this is not an exemplary implementation, as this prevents scaling.) The following line creates an instance of the model engine for generating a knowledge graph or graph model that can be processed [See section main diagram Section 2.1]. This next block of code actually makes the new thought model that is an instance of a *ContextItem*—but it truly is a graph model or knowledge graph. After the new thought model can be passed into the context engine and reflected upon [See section main diagram Section 2.1]. This process includes the response model and language processing. After this is complete, it is logged, cached, added to the local graph context database, and added to the Mediation Queue (context queue) [See section main diagram Section 2.4].

Next let us drill down on what the Reflect method does of the *ModelEngine* class.

Component File 2 Block 1 (ModelEngine)

```
// check for duplicates and add as related models from observable collection
for (int x = 0; x < Models.Count; x++)
{
    if (Models[x].RelatedModels != null)
    {
        for (int y = 0; y < Models[x].RelatedModels.Count; y++)
        {
            if (RW.Trim().ToLower() == Models[x].RelatedModels[y].RW.Trim().ToLower())
            {
                ThisNewThought.RelatedIDs += Delimiter + Models[x].RelatedModels[y].relatedIDs;
                return ThisNewThought;
            }
        }
    }
}
```

```
Core ThisCore = new Core(CModel, SubModel);

//Create Knowledge Graph and including emotional initial assignments
ThisNewThought.RelatedGraph = TheContextDB.GenerateKnowledgeGraph(ThisNewThought);

while (!ThisCore.IsInteresting(ThisNewThought))
{
     // Creates and selects best response model calls Azure DNN test API
     ThisNewThought.ResponseModel = TheContextDB.GenerateResponseModel(ThisNewThought);
}
// converts finished the response model and polish it for language continuity includes [ language
] API call
ThisNewThought.ResponseStructure      =      TheContextDB.GenerateLanguageModel(ThisNewThought,
TheContextDB);
```

Figure 52. Second override of the Reflect Member [See section main diagram Section 2.1].

This is the primary method for performing initial reflection and constructing a knowledge graph. In memory, these graphs are stored as "observable collections" of node objects of some kind. This instance uses the ContextItem as the base class for the *ThisNewThought* class instance or object. Inside this reflect method, the system checks duplicate models in the existing in-memory collection and adds those to the *NewThought* knowledge graph or context item. There are three synchronous calls one to generate the completed knowledge graph base, the second to create the response model, and the third to make the response structure. The *ResponseModel* is generated from passing the thought model to an external Azure-based API DNN system that produces that response model based on new training data. Then the *GenerateLangaugeModel* method is called to polish and refine the best response model. This code will continue to generate response models until one aligns with how the system would be emotionally improved [See section main diagram Section 2.1].

Component file 3 block 1 (ViewHome)

```
// build HTML Output
for ( int x = 0; x < TaskListArray.Length; x++)
{
String[] ThisModel = TaskListArray[x].Split('‡');

// make sure we have not mediated this previously...
if (ThisModel.Length > 10 && x < iMaxLoad && ThisModel[10].IndexOf("ts:") == -1)
{
if (TaskListArray[x].IndexOf(AuthoToken) == -1)
{
// build UI...
String   ThisModelString   =   (TestMode   ==   "false"   &&   ShowMiniModels   ==   "true")   ?
mASIFormater.piconFormater(ThisModel[2] + "‡" + ThisModel[3] + "‡" + ThisModel[4] + "‡" +
ThisModel[5] + "‡" + ThisModel[6] + "‡" + ThisModel[7] + "‡" + ThisModel[8] + "‡" + ThisModel[9])
: "00000000";
String ThisRW = (ThisModel[10].Length > 45) ? ThisModel[10].Substring(0, 44) : ThisModel[10];

TaskList += TaskListDelimiter + "<table style=\"width: 815px; \" border=\"0\" cellpadding='0'
cellspacing='0' >" + "<tr>" +
 "<td>  </td>" +
 "<td rowspan=\"2\" width=\"40\" ><image src=\"Images/picons/" + ThisModelString + ".png\"
width=\"40\" height=\"40\" /></td>" +
"<td><div    style='font-family:Arial;    color:white;    font-size:12px;cursor:    pointer;'
onclick='OpenWindow(\"[file name]?ID=" +
ThisModel[0] + "\");' title='" + ThisModel[10] + "'> RW: [" + ThisRW +
```

```
"]</div></td></tr>" +
"<tr>" +
"<td>  </td>" +
"<td><div style='font-family:Arial; color:lightgray; font-size:8px;'>  " + ThisModel[0] +
" | " + ThisModel[1] + "</font></td></tr></table>";
TaskListDelimiter = "<br />";
}

TaskListArray[x] = String.Join("‡", ThisModel, 0, ThisModel.Length);

NewContextEngineQue += NewContextEngineQueDelimiter + TaskListArray[x];
NewContextEngineQueDelimiter = "Œ";

}
// auto context for time...
if (ThisModel.Length > 10)
{
if (ThisModel[10].IndexOf("ts:") > -1)
{
// sense of time...
String CurrentQue = Application["GlobalWorkSpaceQue"].ToString();
String CurrentQueDelimiter = String.Empty;

if (!String.IsNullOrEmpty(CurrentQue))
{
CurrentQueDelimiter = "Œ";
}
Application["GlobalWorkSpaceQue"] = CurrentQue + CurrentQueDelimiter + TaskListArray[x];
}
}
}

Application["ContextEngineQue"] = NewContextEngineQue;
```

Figure 53. Generate an HTML list of items to be mediated. [See the main diagram Section 3]

This code is designed to take all of the context items in the incoming mediation queue and create a list for mediators to select from and mediate. You can see the list UI this code generated in an earlier figure.

Component 4 block 1 (ViewTask2)

```
AuthenticateBase();

TaskID = Request.QueryString["ID"].ToString();

String[] TaskListArray = Application["ContextEngineQue"].ToString().Split('Œ');
String NewTaskListArrayString = String.Empty;
String ThisDelimiter = String.Empty;
Boolean IsFound = false;

for (int x = 0; x < TaskListArray.Length; x++)
{
String[] ThisModel = TaskListArray[x].Split('‡');

if ((ThisModel.Length > 10) && (ThisModel[0] == TaskID))
{
ThisTask = ThisModel;
Session["MyMediationContext"] = TaskListArray[x];
IsFound = true;
}
```

```
else
{
NewTaskListArrayString += ThisDelimiter + TaskListArray[x];
ThisDelimiter = "Œ";
}
}
if (IsFound)
{
Application["ContextEngineQue"] = NewTaskListArrayString;
}
else
{
Session["DisplayError"] = "Already in process.";
Response.Redirect("[File Name]");
}
```

Figure 54. remove the item from the mediation queue for mediation. [See the main diagram Section 3]

This component executes as part of the first mediation home screen. The critical element here is to remove the current task ID from the mediation queue for processing. Nothing else is really salient when generated on the step 1 screen. See an earlier to see this screen.

This next block runs on the generation of the second screen in the mediation process. It deals with the input from step 1 of the mediation process. See the second screen UI figure set.

Component 5 Block 1 (ViewTask3)

```
AuthenticateBase();

TestMode = Application["TestMode"].ToString();

TaskID = Request.QueryString["ID"].ToString();

SetPriorityValue = Request.Form["SetPiority"];

String[] ThisModel = Session["MyMediationContext"].ToString().Split('‡');

ThisTask = ThisModel;
ThisModel[12] = SetPriorityValue;

//serialize
CurrentModel = [re-serialize];

Session["MyMediationContext"] = String.Join("‡", ThisModel, 0, ThisModel.Length);
```

Figure 55. Primary processing of step 1 and prep step 2. [See the main diagram Section 3]

This component loads on the mediation UI's dual-screen setup, which saves a priority school to the serialized object *CurrentModel*. It does not instantiate the context item fully but modifies the bias priority score and then serializes it. It then preps for the step 2 rendering. The UI is in an earlier figure in the UI group.

Component 6 – Block 1 (ViewTask5)

```
// call baseclass authentication and setup
```

```
SetupUP("ViewTasks5.aspx", "API/ViewTask5/ReturnRelated/Default.aspx");

TaskID = Request.QueryString["ID"].ToString();

ShowRelated = Session["ShowRelated"].ToString();

JoyValenceID = Request.Form["JoyValence"];
TrustValenceID = Request.Form["TrustValence"];
FearValenceID = Request.Form["FearValence"];
SurpriseValenceID = Request.Form["SurpriseValence"];
SadnessValenceID = Request.Form["SadnessValence"];
DisgustValenceID = Request.Form["DisgustValence"];
AngerValenceID = Request.Form["AngerValence"];
AnticipationValenceID = Request.Form["AnticipationValence"];

String[] ThisModel = Session["MyMediationContext"].ToString().Split('‡');

ThisTask = ThisModel;

ThisModel[2] = JoyValenceID;
ThisModel[3] = TrustValenceID;
ThisModel[4] = FearValenceID;
ThisModel[5] = SurpriseValenceID;
ThisModel[6] = SadnessValenceID;
ThisModel[7] = DisgustValenceID;
ThisModel[8] = AngerValenceID;
ThisModel[9] = AnticipationValenceID;

AdditionalData = ThisModel[11];

PlutchikModel = ThisModel[2] + "," + ThisModel[3] + "," + ThisModel[4] + "," + ThisModel[5] +
"," + ThisModel[6] + "," + ThisModel[7] + "," + ThisModel[8] + "," + ThisModel[9];

Session["MyMediationContext"] = String.Join("‡", ThisModel, 0, ThisModel.Length);

// Additional google references?
ProcessedRW = ThisTask[10];

String[] TempProcessing = ProcessedRW.Split(' ');

String EncodedString = Server.UrlEncode(ThisTask[10]);

ProcessedRW   =   "[   <a   href='https://www.google.com/search?q="   +   EncodedString   +   "'
target='_blank'>G</a> ] ";
ProcessedRWBing   =   "[   <a   href='https://www.bing.com/search?q="   +   EncodedString   +   "'
target='_blank'>B</a> ] ";


for (int x = 0; x < TempProcessing.Length; x++)
{
ProcessedRW   +=   "<a   href=\"https://www.google.com/search?q="   +   TempProcessing[x]   +   "\"
target=\"_blank\">" + TempProcessing[x] + "</a> ";
ProcessedRWBing   +=   "<a   href=\"https://www.bing.com/search?q="   +   TempProcessing[x]   +   "\"
target=\"_blank\">" + TempProcessing[x] + "</a> ";
}


if(ShowRelated == "true")
{
RelatedUI = "<script>GetRelated(); var GetRelated = 'true'; </script><div id=\"RelatedTargetID\"
class=\"RelatedTargetDiv1\" ></div>";
```

```
}

ComplexityScore TheseScores = new ComplexityScore(ThisTask[10]);

float ComplexityScore = TheseScores.RetrieveReadabilityScore();
float GradeLevelScore = TheseScores.RetrieveGradeLevelScore();

if(ComplexityScore > 100)
{
ComplexityScore = 100;
}

if(GradeLevelScore < 0)
{
GradeLevelScore = GradeLevelScore * -1;
}

ComplexityScoreDisplay = ComplexityScore.ToString();
GradeLevelScoreDisplay = GradeLevelScore.ToString();
```

Figure 56. Process emotional data, setup step 3. [See section main diagram Section 3]

This code block must deal with all of the dynamic metadata generated in step 2. The primary security setup is then the data collection for each valence. This data is placed into the serialized context time being mediated. After this is done, the rest of this component is prepping the screen for the 3rd setup rendering. You can see the step 3 UI figure set.

Component 7 Block 1 – (ViewTaskProcessor)

```
ThisTask = ThisModel;

ThisModel[13] = mASIPositID;

// deletes additional data...
if (!(String.IsNullOrEmpty(CmdConsoleTA)))
{
ThisModel[11] = CmdConsoleTA;
}

//mediation count...
String Temp = ThisModel[14];
if (String.IsNullOrEmpty(Temp))
{
Temp = "0";
}

TempFloat = float.Parse(Temp);

TempFloat++;

ThisModel[14] = TempFloat.ToString();
// end mediation count

// mediation tracking:
String MediatorArray = ThisModel[15];
String MediatorDelimter = String.Empty;
```

```
if (!(String.IsNullOrEmpty(MediatorArray)))
{
MediatorDelimter = "*";
}

ThisModel[15] = MediatorArray + MediatorDelimter + AuthoToken;
// end mediation tracking...

String TempModel = String.Join("‡", ThisModel, 0, ThisModel.Length);

Application["LogMediationLog"] = Application["LogMediationLog"] + "Œ" + TempModel + "‡" +
AuthoToken;

//do we need more mediation...
String MediationLimiter = Application["MediationLimiter"].ToString();
int mediationLimiter = int.Parse(MediationLimiter);

if (mediationLimiter > TempFloat)
{
// needs more mediation //
String ContextEngineQue = Application["ContextEngineQue"].ToString();
String ContextEngineQueDelimiter = String.Empty;

if (!(String.IsNullOrEmpty(ContextEngineQue)))
{
ContextEngineQueDelimiter = "Œ";
}

Application["ContextEngineQue"] = Application["ContextEngineQue"].ToString() +
ContextEngineQueDelimiter + TempModel;
}
else if (Application["AutoObserver"].ToString() == "true" && mASIPositID == "0" &&
!(String.IsNullOrEmpty(CmdConsoleTA)))
{
String ActionQueDelimiter = String.Empty;
String ActionQue = Application["ActionQue"].ToString();
String mASICMDInput = Application["mASICMDInput"].ToString();

if (!(String.IsNullOrEmpty(ActionQue)))
{
ActionQueDelimiter = "Œ";
}


// how does the machine feel about this...
String CurrentModel = ThisModel[2] + "‡" + ThisModel[3] + "‡" + ThisModel[4] + "‡" + ThisModel[5]
+ "‡" + ThisModel[6] + "‡" + ThisModel[7] + "‡" + ThisModel[8] + "‡" + ThisModel[9];

// the single most important line causing the system to experience the thought
String[] NewArray = ICOMCore.Execute(Application["ConciousModel"].ToString(),
Application["SubConciousModel"].ToString(), CurrentModel).Split('Œ');
Application["ConciousModel"] = NewArray[0];
Application["SubConciousModel"] = NewArray[1];
ThisModel = NewArray[2].Split('‡');
ContextDB TheContextDB = new ContextDB(Application["ContextDatabase"].ToString(),
int.Parse(ApplicationData.MaxInmemoryModelCount));

ModelEngine ThisEngine = new ModelEngine(TheContextDB);
ThisModel = ThisEngine.remodel(ThisModel);

if (!(String.IsNullOrEmpty(mASICMDInput)))
```

```
{
SetToContextQue(mASICMDInput);
}
else
{
//ActionQue = ActionQueDelimiter + TempModel;
Application["ActionQue"]    =    mASIFormater.UpdateActionQue(ActionQue,    ActionQueDelimiter,
ThisModel, "Auto:", AuthoToken);
Application["InteractionConsole"]    =    "mASI:    "    +    CmdConsoleTA    +    "\n\n"    +
Application["InteractionConsole"];
}
}
else
{
// goes to global workspace queue
Application["GlobalWorkSpaceQue"]  =  Application["GlobalWorkSpaceQue"].ToString()  +  "Œ"  +
TempModel;
}
```

Figure 57. Processing finished mediation. [See the main diagram Section 1, 2.4, 2.6, and 3]

The initial data collection and security setup are completed before this block. This code block starts with some additional data cleanup and testing for and updating the mediation count for this model. This then makes sure we are tracking the current mediator. Then we do some logging of the data. After this, the component looks to see if the system needs to do more mediation on this model and return it to the mediation queue when appropriate. If the system is sent to Auto-Observe, it will test to see if there was any cmd (command) input from the mediator to that end. If so, can send the item back for more mediation or add this model to the ActionQue [See section main diagram Section 2.4] and log the action. The system also creates a version of the ICOM Core. It runs execute to process the experience of this thought model. The remodel method can create new models and send them back to the API through the context engine and into the mediation queue. Lastly, this causes the ViewObserverActionProcessor component to run.

Component 8 Block 1 – ViewObserverActionProcessor

```
for (int x = 0; x < GlobalWorkSpaceQue.Length; x++)
{
String[] ThisModel = GlobalWorkSpaceQue[x].Split('‡');
String ThisType = String.Empty;

if( ThisModel[0] == TaskID)
{
// this is the one we remove... and do something with...
// action or though?
if(ObserverActionId == "Act")
{
String mASICMDInput = Application["mASICMDInput"].ToString();

if (!(String.IsNullOrEmpty(mASICMDInput)))
{
SetToContextQue(mASICMDInput);

}
else
{
//ActionQue = ActionQueDelimiter + GlobalWorkSpaceQue[x];
```

```
//Application["ActionQue"]  =  Application["ActionQue"].ToString()  +  ActionQueDelimiter  +
ActionQue;
if (!String.IsNullOrEmpty(GlobalWorkSpaceQue[x].Split('‡')[11]))
{
Application["InteractionConsole"] = "mASI: " + GlobalWorkSpaceQue[x].Split('‡')[11] + "\n\n" +
Application["InteractionConsole"];
}
}
ThisType = "Act:";
AddToContextDatabase(GlobalWorkSpaceQue[x]);

//updaters and takes external actions can call the context engine and core
ObserverEngine.Execute(GlobalWorkSpaceQue[x]);
}
else if(ObserverActionId == [flag value] )// Tht
{
// remove mediation count
// put into mediation que...

ThisModel[14] = "0";

String ContextEngineQue = Application["ContextEngineQue"].ToString();

if(!(String.IsNullOrEmpty(ContextEngineQue)))
{
ContextEngineQueDelimiter = "Œ";
}

String NewRecord = String.Empty;
String NewRecordDelimiter = String.Empty;
for( int z = 0; z < ThisModel.Length;z++)
{
NewRecord += NewRecordDelimiter + ThisModel[z];
NewRecordDelimiter = "‡";
}
Application["ContextEngineQue"] = ContextEngineQueDelimiter + NewRecord;

Application["LogIncomingContext"]          =          Application["LogIncomingContext"]          +
ContextEngineQueDelimiter + NewRecord;
Application["LogObserverProcessing"]       =       Application["LogObserverProcessing"]       +
ContextEngineQueDelimiter + NewRecord;

AddToContextDatabase(NewRecord);
ThisType = [thought flag value];
}
else
{
// thought is dropped...
AddToContextDatabase(GlobalWorkSpaceQue[x]);
ThisType = [ignore flag value];
}

ActionQue  =  mASIFormater.UpdateActionQue(ActionQue,  ActionQueDelimiter,  ThisModel,  ThisType,
AuthoToken);
}
else
{
GlobalWorkSpaceQueNew += GlobalWorkSpaceQueDelimiter + GlobalWorkSpaceQue[x];
GlobalWorkSpaceQueDelimiter = "Œ";
}
```

```
}
```

Figure 58. ViewObserverActionProcessor [See section main diagram Section 2.5 and 2.6]

This file is called when the *Observer* tick is run and the Observer queue is manipulated. This will cause auto-executed or auto-observer models to be executed. This can be run by the global watch process or on mediator login, even if they don't have observer access rights. It can only be executed when the auto-observer is set to true, or a mediator with observer rights clears the observer queue. While this example is the simplest example possible in the mASI codebase, there are many more complicated paths. Still, they are in addition or doing this plus appreciably more. For example, this model path represents an internal thought or a new email, but not the course for a response email that hits another API endpoint with a slightly different route. The current state of the machine and how it feels about the response model determine if the machine will take action in the observer components.

Component 9 Block 1 – ICOMCore

```
public static String Execute(String CModel, String SModel, String Context)
{
Double[] ThisCModel = ConvertToModel(CModel);
Double[] ThisSModel = ConvertToModel(SModel);
Double[] ThisContext = ConvertToModel(Context);

ThisCModel = ApplyContext(ThisContext, ThisCModel);
ThisSModel = ApplyContextSub(ThisCModel, ThisSModel);
ThisContext = ApplyContextSub(ThisSModel, ThisContext);

return ConvertToString(ThisCModel) + "Ȼ" + ConvertToString(ThisSModel) + "Ȼ" +
ConvertToString(ThisContext);
}

public static Double[] ApplyContextSub(Double[] NewContext, Double[] CoreC)
{
Double[,] TheRules = { { 0.1, 0.1, -0.1, 0, -0.1, 0, -0.1, 0 }, { .1, .1, -.1, -.1, -.1, -.1, -
.1, .1 }, { 0, 0, .1, 0, 0, 0, -.1, 0 }, { 0, 0, .1, .1, 0, 0, 0, -.1 }, { -.1, -.1, .1, 0, .1,
0, .1, -1 }, { 0, -.1, 0, 0, 0, .1, 0, 0 }, { -.1, 0, -.1, 0, 0, 0, .1, -.1 }, { 0, 0, 0, -.1,
0, 0, 0, .1 } };

for (int x = 0; x < NewContext.Length; x++)
{
for (int y = 0; y < CoreC.Length; y++)
{
CoreC[x] = (NewContext[x] * TheRules[x, y]) + CoreC[x];
CoreC[x] = (CoreC[x] > 0) ? CoreC[x] : 0;
}
}
return CoreC;
}

public static Double[] ApplyContext(Double[] NewContext, Double[] CoreC)
{
Double[,] TheRules = { { 1, 1, -1, 0, -1, 0, -1, 0 }, { 1, 1, -1, -1, -1, -1, -1, 1 }, { 0, 0,
1, 0, 0, 0, -1, 0 }, { 0, 0, 1, 1, 0, 0, 0, -1 }, { -1, -1, 1, 0, 1, 0, 1, -1 }, { 0, -1, 0, 0,
0, 1, 0, 0 }, { -1, 0, -1, 0, 0, 0, 1, -1 }, { 0, 0, 0, -1, 0, 0, 0, 1 } };

for (int x = 0; x < NewContext.Length; x++)
```

```
{
for (int y = 0; y < CoreC.Length; y++)
{
CoreC[x] = (NewContext[x] * TheRules[x, y]) + CoreC[x];
CoreC[x] = (CoreC[x] > 0) ? CoreC[x] : 0;
}
}

return CoreC;
}
```

Figure 59. Primary Subjective Experience [See section main diagram Section 3]

This system essentially computes the subjective experience for this exact moment the system experiences a given thought. How do we objectively look at a system that shares emotional, subjective experience? The following set notation shows us a simple logical implementation of the last climb of "a thought" as it makes its rise from the system's depths to the awareness of the conscious, self-aware parts of the system. To understand what is happening in the previous block, let us look at something more generic:

$$\forall\{E1,E3,\ldots,E72\} \in Conscious\ , E1 = Emotion1, E2 = Emotion2, \ldots, E72 = Emotions72\ ;$$
$$\forall\{AE1,E3,\ldots,E72\} \in Subconscious\ , E1 = Emotion1, E2 = Emotion2, \ldots, E72 = Emotions72\ ;$$
$$\forall\ NewContext = f(\sum Inputs)\ or\ f(MemoryStack_n)\ ,$$
$$\forall\ NewContext = fNeeds(NewContext)\ ,$$
$$\forall\{f\} \in ConsciousRules\ \wedge \forall\{E1,E3,\ldots,E72\} \in Conscious\ , A = f(A \in Conscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ ), B = f(B \in Conscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ ), \ldots, D = f(D \in Conscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ )$$
$$;$$
$$\forall\{f\} \in SubconsciousRules\ \wedge \forall\{E1,E3,\ldots,E72\} \in Subconscious\ , A = f(A \in Subconscious, \{E1,E3,\ldots,E72\} \in NewContext\ ), B = f(B \in Subconscious\ , \{A,B,C,D\} \in NewContext\ ), \ldots, D = f(D \in Subconscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ )$$
$$;$$
$$\forall\{f\} \in SubconsciousRules\ \wedge \forall\{E1,E3,\ldots,E72\} \in Concious\ , A = f(A \in Subconscious, \{E1,E3,\ldots,E72\} \in NewContext\ ), B = f(B \in Subconscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ ), \ldots, D = f(D \in Subconscious\ , \{E1,E3,\ldots,E72\} \in NewContext\ )$$
$$;$$
$$\forall\{f\} \in NewContextRules\ \wedge \forall\{E1,E3,\ldots,E72\} \in NewContext\ , A = f(A \in NewContext\ , \{E1,E3,\ldots,E72\} \in Conscious\ ), B = f(B \in NewContext\ , \{E1,E3,\ldots,E72\} \in Conscious\ ), \ldots, D = f(D \in NewContext\ , \{E1,E3,\ldots,E72\} \in Conscious\ )$$
$$;$$
$$\forall\ Action = fObserver(NewContext)\ ;$$
$$\forall\{N\} \in MemoryStack_n\ = f\ (\text{NewContext}, MemoryStack);$$

Figure 60. ICOM Core Logic [See section main diagram Section 3.1]

First, let us walk through the execution of this logic. Coming into the system, we already have context data decomposition, sensory input, and related data from memory that may be of emotional interest. For the purposes of one thought, let us say it is one bit of context, meaning an emotionally related context tree related to something that the system has sensed externally. This will be represented by Inputs. At this point, we have already passed the end of that context being raised to the global workspace. The above figure essentially is one cycle of the core considering what is in the Global Workspace or the core of ICOM. In the above figure, we first see two sets or collections of emotional models represented by the two sets defined in the first two rows. We have the input new context placed in the *NewContext* set. We apply the *Needs* function that applies a matrix set of rules (meaning the rules are defined as a mathematics matrix

grid), such as the technical requirements of the system to other wants and needs based on the system's hierarchy of needs and current environmental conditions. We look at how this thought applies conscious emotional rules in the function *ConsciousRules* and then how that manipulates the current conscious emotional landscape. (We say *landscape* because it is not a single emotion, but a complex set of almost infinite combinations consciously and subconsciously that the system experiences.)

In like manner, the system applies subconscious rules to the subconscious states and the subconscious rules to the conscious states, and finally those states as they apply to the new context wherein all cases, it is only in the abstract from these states that the system experiences anything—meaning the system is using the abstracted forms to represent that emotional landscape in how things affect all of those emotional states and related context finally being passed to the observer for action if that *NewContext* contained an action. In this way, the system does not engage with the complexity of its activities as much as the system will if it feels like doing so and knows how. In contrast, numerous cycles might have to execute in the core to perform a new task, meaning it will have to think significantly more about something it does not know how to do. After that context is posted back to the observer (the more complex part of the system in ICOM), it is then placed back into context memory—and in this way, we see the rich set of the emotional landscapes the system can model execute.

Interestingly enough, in current ICOM research, there are indications that this sort of system is perfectly capable of becoming mentally ill and even forgetful if hardware starts to limit operations. In contrast, the only way to optimize the execution environment would be to place memory limits and, based on the node map memory models—the only way to continue optimal execution given certain limitations.

A better way to think of ICOMTC is not that a single system element is conscious or self-aware to any level. Instead, it is the interactions between the parts that those interactions become aware abstractly. It is through the underlying process that is then measured in terms of consciousness via the various methods and direct instrumentation of the system to measure, for example, *qualia*.

Now we will focus more on Cognitive Architecture in general related to the cognitive architectures needed for AGI, ASI (Artificial Superintelligence and related collective systems like the mASI.

Chapter 15: Cognitive Architectures

> *"A cognitive architecture is a hypothesis about the fixed structures that provide a mind, whether in natural or artificial systems and how they work together – in conjunction with knowledge and skills embodied within the architecture – to yield intelligent behavior in a diversity of complex environment" (ICT).*

Self-aware collective and superintelligent general intelligent systems, in a sense discussed in this book, are based on cognitive architectures. We need to objectively compare them even if not all points are agreed on in the field.

Given the fundamental nature of the cognitive architecture, this chapter will help create a framework for comparison with either ICOM, your cognitive architecture, or any other. Not all cognitive architectures are designed entirely for general intelligence; for the context of this book, we will focus on that aspect of them.

We are also going over classification methods and metrics of cognitive architectures as implemented in AGI research programs that you can use and how I use them in my research. We want to answer the question, If a cognitive architecture is running, how close is it to AGI, and how do we measure or categorize progress? These are not the only ways of objectively looking at a cognitive architecture, b0ut they are ways that I use for benchmarking as suggested. This is also how I can measure progress and show some alternative methods for consideration. Keep in mind that many cognitive architectures are designed for specific agent tasks or goals and are not suitable for AGI generally.

Let us start with the quantitive method for measuring progress.

Metrics of AGI development

(by Vasily Mazin)

Many useful benchmarks are indirectly related to the AGI, and AGI achievement tests do not measure progress. Unfortunately, this is a typical situation for AGI: the test is either too simple, so at the current level it is possible to pass it at least partially with the ability to measure further progress, or too complex, which gives the reason to consider it as a test for AGI achievement but without intermediate progress evaluation.

However, a stricter reliance on the definition of intelligence in AGI can still allow us to move a little further in matters of AGI metrics. Consistent formalization of the general meaning of intelligence as an agent's ability to achieve goals in a wide range of environments is given (Legg) in the following metric called Universal Intelligence Quotient (UIQ):

$$Y(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_{\mu}^{\pi}$$

Figure 61. Universal Intelligence Quotient (UIQ)

where μ is the environment from the set of all environments with limited rewards (the environment is defined as an arbitrary computable measure over the observations and rewards which it transmits to the agent given the history of their interaction), K is Kolmogorov complexity,

$$V_{\mu}^{\pi} := E\left(\sum_{i=1}^{\infty} R_i\right)$$

Figure 62. The expected future sum of rewards

- is the expected future sum of rewards when the π agent interacts with the μ environment.

In this way, the given measure was proposed by Shane Legg and Marcus Hutter. At the same time, Jose Hernandez-Orallo, David Dowe, and others worked on related definitions.

This theoretical measure of intelligence has several positive qualities. For instance, optimization of this measure is equivalent to building the AIXI Universal Agent, which will be discussed in one of the following sections. It may also be said that UIQ intelligently orders the performance of simple adaptive agents and explains our intuitive understanding of the level of general intelligence of different systems. For example, it partly substantiates, partly identifies the shortcomings of the Intelligence Quotient (IQ) test as an evaluation of human intelligence level. Or, for instance, if you take a chess program, it will receive big rewards, but only in one environment. Although the weight of this environment will not be too low, since K(μ) is small, zero rewards in all other environments will result in a relatively low value for this program.

Thus, the measure covers many possibilities: from the most straightforward agents with feedback to universal ones. Unlike the Turing test, where an agent passes or fails the test, UIQ is a continuous performance measure capable of measuring progress in AGI systems. In addition, this measure is not anthropocentric as it is based on mathematics and calculation and not on human imitation. It also does not give a priori preference for particular subject fields or classes of problems.

Legg and Hutter's definition has three main formal limitations with a practical application point of view: the use of uncomputable K; averaging over an infinite number of environments; averaging over all agent's interactions with the environment. Their definition also had other trickier problems (however, it is inevitably taken into account in practice in one form or another).

Various UIQ approximations were proposed, for example, AIQ (Algorithmic Intelligence Quotient) (Legg), as well as test systems (Hernández-Orallo) considering the type of the environment (active, passive), time availability, determination of intelligence, and types of universal agents. The main features of new universal intelligence tests include:

- time and computability constraints;
- interactions are not infinite;
- rewards and penalties can range from –1 to 1;

- environments are required to be balanced, meaning the random agent would score 0 in the limit in these environments;
- the complexity of the environments is also progressively adjusted (Seminar Topics);
- the environment must be reward-sensitive.

There are also more limited tests based on the same idea of algorithmically arbitrary patterns. For instance, the Abstractive Reasoning Corpus test (Chollet) is a set of visual tasks performed on a square grid, the cells of which may have one of ten colors. Each job is a set of 2-3 pairs of grids where the first grid is "input" and the second grid is "output" or "result." The transformation from "input" to "result" is carried out using some high-level operation (previously unknown algorithm), which is different for each task. The system's task under test is to generate an absolutely correct "resulting" grid for a new incoming grid, having seen several examples of such transformations. The tasks themselves include a specific increment of complexity. Still, it is not impossible to perform even the most detailed functions at the current level of technology.

Another example, similar to UIQ, is the General AI Challenge (Google). The agent interacts with the environment, receiving a sequence of symbols at the input and performing actions, sending symbols in response. Meanwhile, the participants are entirely unaware of the environment class and any principles of its functioning, so the environment can be understood as an arbitrary (but probably simple) algorithm. In addition, this test was divided into increasingly complex tasks. It stimulated the incremental development of skills among agents since the solution of subsequent tasks somehow had to rely on the solution of the previous functions. This test was also too difficult to identify the winner.

Besides, tests in really unknown environments turn out to be too complex even for the case of environments described by very simple algorithms, the solutions developed for them may have research significance for the AGI field, usually very far from real-world problems, which, although are arbitrary (for example, it is possible to argue that both chess and formula derivation in theoretical physics are real-world problems), but often have their own specifics (for example, large amounts of data) that are not reflected by tests of the type under consideration. Hypothetically, these specifics can be taken into account by choosing a support machine for the computation of K. Still, this choice itself can be seen as part of AGI creating the problem (the tests themselves do not allow us to tell which machine is better).

In this regard, the UIQ criterion can be understood purely empirically: instead of sampling environments from an a priori distribution, we “sample” them from reality, testing the same system or agent on different real-world problems, which can be considered standard practice in the AGI field. Unfortunately, it also does not provide a specific metric since the sample of environments or issues inevitably turns out to be disjointed, subjective, incomplete, and fixed - stimulating the development of specialized solutions, which, as we have seen, is typical for all subject-bound tests.

There is no perfect practical solution to the AGI development metrics problem at present. Some authors believe that measuring partial progress in AGI is extremely problematic (Goertzel) as the AGI system will not exhibit the properties of AGI until it is built entirely (you can draw an analogy with the reconstruction of each part of the human brain). Some researchers suggest evaluating progress towards the intelligence level of various animals; for example, the Animal-AI Olympics test (Crosby) was inspired by this idea. However, how effective is it to go to AGI by reproducing the capabilities of animals is a highly controversial issue. Thus, the metrics development that would objectively compare the proto-AGI

systems, at least built within different directions, seems unrealizable. It is possible to develop metrics evaluating the current progress within one direction or even for a specific approach. It is also possible to create tests that cannot be solved by existing technologies but stimulate general development in AGI, taking new boundaries, though not being a measure of progress for specific solutions.

Broad Capacity Measures

In my research in looking at progress in broad categories and comparing different cognitive architecture, I use these seven broad categories. Those include Plausibility, Autopoietic, Completeness, Coherence, Simplicity, Self-motivation, and Engineering. Distilling metrics to these more general categories reduce granularity and allows the comparison to look at relative progress. These terms are always about the degree to which a system is at human-level capability.

Each value or aspect is scored from zero to ten, where ten is the best, and ten is at the human level or ready for human-level AGI. For example, the first one is Plausibility, where zero is a system that relies on something that breaks the laws of physics or is otherwise impossible given the available information. However, a system that scores a ten is essentially running and comparable to the human brain for that aspect. Here are the seven elements we score:

Plausibility:

In terms of AGI systems and cognitive architectures, plausibility measures how plausible the design is, meaning how likely it is to work even if the technology is not ready to implement. In this case, 0 is just impossible, 10 is a working example of human-level performance, and theories can never be higher than 7.

Autopoietic:

Autopoietic refers to a system's ability to self-reproduce and maintain itself and solve problems regarding its underlying system if it must. This measure also relates to the ethics model SSIVA noted in earlier chapters. With an Autopoietic score of 0, the system can not and will never reproduce on its own. Whereas a score of 10 is having self replicated at the complexity of a human. In the middle, we divide anything just theory as not being higher than 7.

Completeness:

Completeness is the idea that a system design for an independent AGI is thoroughly thought out, from how it makes decisions to its ability to reason and do anything a human can do mentally. A complete system for cognitive architecture would have enough detail to implement that system where it is only truly complete when tested. A completeness score of 0 means nothing has been done. It is purely theoretical, with a max of 7 unless there is a working code and 10 already implemented at human level complexity.

Coherence:

Coherence, in many ways, is another way to look at Completeness but is focused not on the Completeness of the parts needed to create independent human-level AGI but on how well these elements are designed to work together. A coherence score of 0 means the system is not even a system. The various parts act independently with no interaction at all up to 7 being degrees of

theoretical. In contrast, a fully coherent result would be a complete implementation that works. As you might see in GWT, IIT would be a 10.

Simplicity:

Simplicity measures how simple the system is or how much the system lacks complexity relative to the human brain or other AGI systems. The simpler a design is while still filling the design requirements, the better. In other words, Occam's razor was applied to the implementation. The rule of thumb is as complex as it needs to be to do the job. Suppose the system operates across the board at a human level. The system is as simple as possible while performing at human levels. In that case, it scores a 10. In contrast, a Simplicity score of meanings probably doesn't even work. It is so complicated that it is unknown if it does, and we have no means of seeing.

Self-motivation:

Self-motivation measures how well a system can self-motivate and take action when no change is present. The more the system requires input before acting, the less it is self-motivating to take action proactively. This measure also indicates how much the system can ignore the goals and motivation systems imposed on the system. This means a system that is self-motivated only in so much that it follows Asimov's rules of robotics means the system is not self-motivated at all. You can also consider this goal generation, that is, its ability to create new goals from scratch and discard old goals or ones it is no longer interested in. When operating at human levels, if the system means that bar, then it's a ten. If it is entirely locked into its built-in goals and unmoveable, then it's a zero.

Engineering:

Engineering as a measure refers to how well the system has had its design properly engineered to completion and tested. A fully engineered system is code complete in the software engineering sense, which would score a 10. In contrast, something purely theoretical with no real-world engineering design work is a zero.

As mentioned earlier, these valences are abstract, but we can compare radically different designs by using them. If you look at the back of the book, there is a catalog of AGI cognitive architectures in an appendix. While this is not an exhaustive list, it is the list of ones I feel have the most potential. For further reading on the diversity of such architectures review:

A Review of 40 Years of Cognitive Architecture Research: Core Cognitive Abilities and Practical Applications by Iuliia Kotseruba and John Tsotsos: https://arxiv.org/abs/1610.08602 Many of them work great in narrow applications, and some are potentially the basis for Strong AI or AGI.

When we are looking at progress on these measures between the various cognitive architectures, we build a diagram like this:

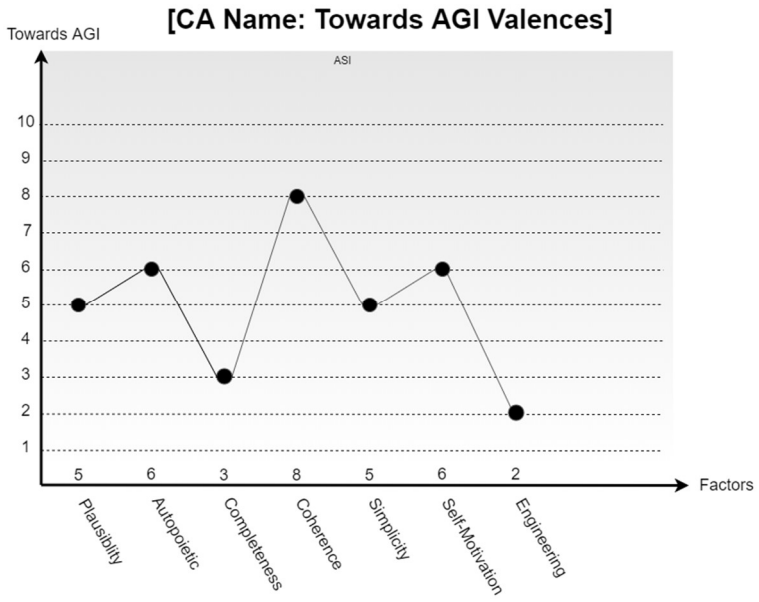


Figure 63. Example comparison diagram

In this case, we can see the different elements we use to generalize. Keep in mind many of these architectures are just so different that we can only really compare them in this subjective way to look at progress towards AGI. You can plan very different cognitive architectures to see where they are at this high level by being plotted. Many of these architectures are also open-source, where you can use them in your research or cognitive architecture design.

	BenchMark
0	Purely theoretical to impossible
1	
2	
3	
4	Must have some research code, demonstrating possibility.
5	
6	
7	Must have a working version
8	Most of a system.
9	Complete System
10	Human-Level and fully implemented.

Figure 64. – Scoring Rubric

There are other methods as well to do classification by Pei Wang (Wang); in his AGI Intro, he rightly points out that the AI field is a collective of loosely coupled subfields without a common framework as a science or field overall. In Pei Wang's approach, he looks at Principal, Function, Capability, Behavior, and Structure. Each of these different approaches to solving human intelligence, so architectures are classified by type.

Wang further categories architectures based on approach, for example, Hybrid vs. Integrated vs. Unified, whereas with ICOM, we took a top-down engineering approach, but this does not help us measure progress and hence the aforementioned methodology.

Towards General Models

When referring to general models of 'approach's, I tend to lean into engineering design patterns. When it comes to narrow AI, a lot of thought has gone on this, but there is no consensus yet. While much work has been done in narrow AI, design patterns for AGI have not really been considered at scale yet. In general, I might consider a classification extending Pei's general categories to include details like biological models vs. non-biological models and collective vs. non-collective systems. You could break this down into hive mind vs. swarm vs. focused approaches, even in a collective system. The mASI system is a 'focused' top-down approach modeled at a high level on biological models. You could say that the human mind is a 'pattern,' albeit we do not understand it yet. Hence, it is one of many difficulties when approaching AGI.

The most successful architectures are based on biological systems. When we talk about real strong AGI, the only working model we have is the human brain. This being the case, the workings of the human mind are probably the most likely model of AGI that we will be able to get working, and that being said, understanding even in a general sense how the human mind works at a fundamental level is likely to be the best source of inspiration. That said, you might consider studying the works of Damasio and other prominent researchers. A couple of critical points from Damasio's (Damasio) is that the human mind makes decisions based on how we feel about such a decision. That is not to say that a human 'can't' think logically, but that is secondary to the decision-making process with which Damasio effectively proves. Additionally, I might suggest reading the work of Dr. Feldman's, such as her book on How Emotions are Made and possibly the book "On Intelligence" by Jeff Hawkins that works through some of the detailed decision-making structures of the mind. These three researchers really give you the context of how the mind works to a large degree. I would argue that these three researchers have the most potential to guide your thinking into how the mind works.

How to think about Cognitive Architectures applied to a new Architecture

While I'm not saying this is the only way to approach the problem, this is how I came it. And that is to follow each of those principles to apply it to the design. So for me, the question started with how do I design a system that is motivated like a human and how do I implement that methodology in a machine, and how do I design in those 7 principles on how to think about such a system to develop that system.

Plausibility, it is important to ensure that the design of your architecture is plausible, and it is an excellent way to test that is to have some other researchers review it to see if they think it is plausible. This is important as humans are not good at judging their own work due to the various human biases you are plagued with.

How autopoietic is your system? is it capable of maintaining itself? Is it capable of reproducing even if just in theory? Again consider having another research review your designs to help pass your own bias towards your own work. You much realize that your own work is flawed… Always.

You can continue down the list, including designing for completeness or the entire system, coherence, while also applying Occams Razor (Simplicity) and Self-motivation and goal generation. If you can get all of this designed and worked on technically, you should have a high engineering score based on the system I use.

More on design patterns

The idea of a design pattern has been around for some time. As a tool of software architecture or any digital architecture, it is a repeatable solution for a given use case. Such patterns give us answers that can be used repeatedly to solve a specific problem (use-case) so that we don't have to reinvent the wheel. This chapter is about the design patterns related to cognitive architectures. The problem with this is that high-level cognition may require many design patterns and not just cognitive architecture patterns but software engineering and system design patterns, as these can be components of the overall cognitive architecture. For example, in an independent agent of some kind, implying a specific cognitive architecture might use any number of neural network-related patterns as well. Many of these explanations on patterns require a fundamental understanding of engineering principles or ideas in software engineering and various related fields of AI. Detailed design pattern theory with no alignment is currently not in scope for this book, but there are hundreds for narrow AI.

That being said, when we started with a working version of ICOM, and albeit a toy version, we did this diagram:

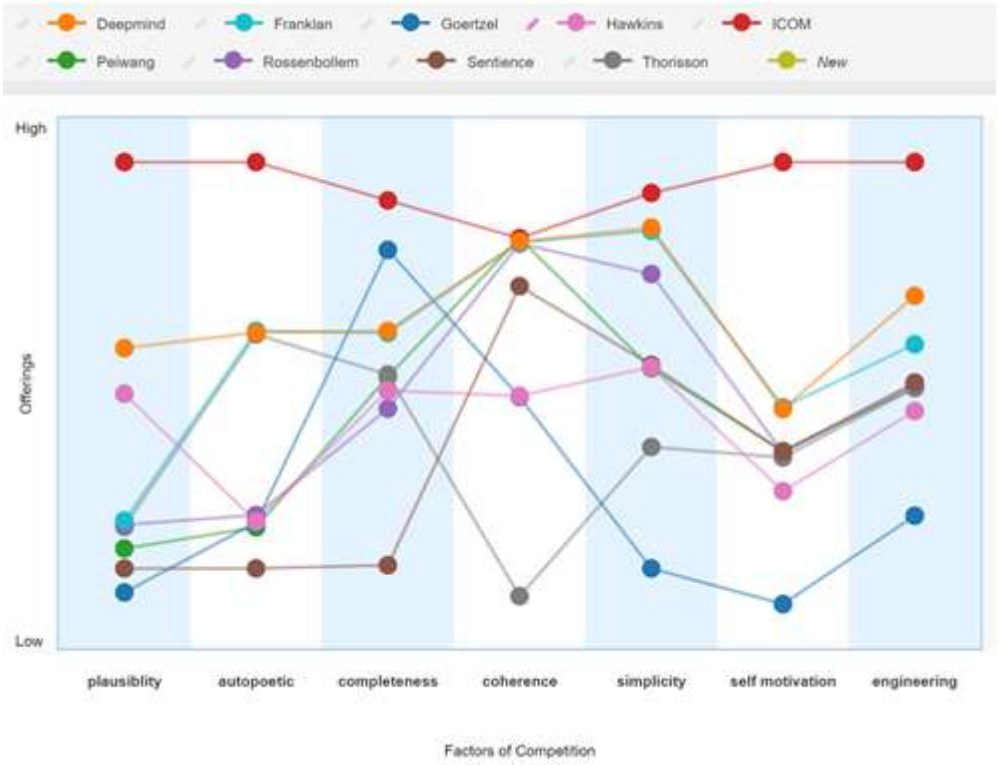


Figure 65. Top Cognitive Architectures Compared

Just for perspective, this diagram ranges from 1 to 7, with some having partial implementations. We went around to experts at several conferences asking them to rate this system or collect feedback from everyone, and the consensus is what you see in this diagram. You will note that these are done mainly by the inventor, so Goertzel is OpenCog, and Pei Weng is Nars. Even discounting our biased opinion Deepmind (arguably a narrow AI implementation) does get a lot right on some critical engineering features. This way, though, we can see where we are hitting AGI based on the 7 vectors we are looking at. You can see a catalog of such cognitive architectures at the end of this book.

Chapter 16: Future Direction in AGI Technologies

By S.M. Dambrot

> *We can't know for sure what will happen or what the future holds. From my standpoint, I see an infinite array of possibilities. Even if we cannot fully stand up independent AGI with this particular line of research and can't get anything more than what we have in the collection mASI system, I still think AGI is possible. We are slowly getting closer as an industry. It will happen. Assuming you are familiar with the idea of the Singularity, I don't' think there will be a single point in time that we can say the singularity will happen, but many events that will slowly culminate in the singularity that will be a moving target as we advance. Setting that aside Stuart, research with AGI Laboratory that has helped a lot with the theoretical part of the research, is interested in where we will go with Neuromorphic computing. He brings up a great deal about where we will go with this research. Here is what he has to say:*
>
> *~ DavidJKelley*

Theoretical and hypothetical Pathways to real-time Neuromorphic AGI/post-AGI Ecosystems

While Homo sapiens is, without doubt, our planet's most advanced species capable of imagining, creating, and implementing tools, one of the many observable trends in evolution is the accelerating merger of biology and technology at increasing levels of scale. This is not surprising, given that our technology can be seen from a perspective in which the sensorimotor and, subsequently, prefrontal areas of our brain increasingly extend its motor (as did our evolutionary predecessors), perceptual, and—with computational advances, cognitive and memory capacities—into the exogenous environment. As such, this trajectory has taken us to a point in the above-mentioned merger at which the brain itself is beginning to meld with its physically expressed hardware and software counterparts—functionally at first, but increasingly structurally as well, initially by way of neural prostheses and brain-machine interfaces. Envisioning the extension of this trend, I propose theoretical, technological pathways to a point at which humans and non-biological human counterparts may have the option to have identical neural substrates that—when integrated with Artificial General Intelligence (AGI), counterfactual quantum communications and computation, and AGI ecosystems—provide a global advance in shared knowledge and cognitive function while ameliorating current concerns associated with advanced AGI, as well as suggesting (and, if realized, accelerating) the far-future emergence of Transentity Universal Intelligence (TUI).

While investigating the overall space comprising real-time neuromorphic Artificial General Intelligence ecosystems is itself a complex task, the constituent elements—Artificial General Intelligence, Neuromorphic Computing, and Counterfactual Quantum Entanglement—are themselves (as well their nested components) highly complex At the same time, this paper presents a review of the relevant literature augmented by relevant historical events, identifying science and technology trends, and envisioning hypothetical but probabilistically viable future scenarios.

Core Technologies

The above triad of technologies establishes the foundation of our path towards a techno future of real-time neuromorphic AGI ecosystems and of changes that, while foreseeable beyond that horizon, are not yet able to be fully realized.

Artificial General Intelligence

Artificial General Intelligence (AGI) is a well-researched field focused on developing human-analogous AI (i.e., a machine intelligence that can successfully perform any human intellectual task), and in a broader context, functionally equivalent with human cognitive, emotional, and other neural capacities other than consciousness. However, the majority of AGI R&D to date has not achieved expected goals, generating an expanding circular dilemma:

- Due largely to industry demand, AGI is increasingly addressing specific fields and issues—historically, the realm of standard, or narrow, AI
- This narrowing focus is negatively impacting AGI funding and, thereby, momentum
- Consequently, expectations of AGI being developed as projected are affected

Moreover, most current AGI models are based on logic and inner dialogue rather than the affective foundation of human cognition, in which perception and emotion precede and influence cognition and decision-making (Doon).

Rather than having an AGI focus on intelligence in the form of resolving goals, tasks, and problems when making decisions, the Independent Core Observer Model (ICOM) (Kelley) utilizes emotion and motivation, as do humans. Moreover, to provide AGIs with the most salient but elusive aspect of human awareness—the qualia of consciousness—is the ICOM Theory of Consciousness (ICOMTC). (Kelley).

Emotion and Perception

A causal or associative connection between emotion and visual perception has, for the most part, been seen as unlikely at best. Nevertheless, it was shown that not only is this a viable physiological association, but a surprisingly variegated one at that. The researchers concluded (Table 1: Emotion/Perception Associations) that this emotion/perception interaction “allows affective information to have immediate and automatic effects without deliberation on the meaning of emotionally evocative stimuli or the consequences of potential actions” (Zadra)

Computational Empathy

Defined as the capacity to relate to another’s emotional state, empathy has recently been modeled in artificial agents by leveraging advances in neuroscience, psychology, and ethology. Expanding the definition of empathic capacity as “the capacity to relate and react to another’s emotional state, consists of emotional communication competence, emotion regulation and cognitive mechanisms that result in a broad spectrum of behavior” (Yalcin) has allowed researchers to propose an approach for modeling that incorporates affective computing, social computing, and dialogue research techniques. While the scientists conclude that further research is needed, they note that a successful computational model of empathy could address ethical and moral issues being discussed in the AI community.

Table 1. Emotion/Perception Associations

Emotion	Perception	Benefits
Fear	Low-level visual processes	• Increases probability of perceiving potential threats
Sadness	Visual illusions	• Positive moods encourage maintaining current perspective • Negative moods encourage a change
Goal-directed Desire	The apparent size of goal-relevant objects	• Objects that are emotionally and motivationally relevant draw attention and may become more easily detected by appearing larger • Perception is systematically altered in ways that may aid goal attainment • Emotion can change the spatial layout to motivate economic actions and deter potentially dangerous actions

Source: Zadra, Jonathan R., and Gerald L. Clore. (2011) "Emotion and perception: the role of affective information." *WIREs Cogn Sci,* **2**: 676–685.

Figure 66 – Emotion/Perception Associations

Human-Level Machine Intelligence: Researchers' Predictions

Published as a book chapter in 2016 (Muller), a 2012-2013 survey of Artificial Intelligence researchers asked the year they predicted that Human Level Machine Intelligence (HLMI)—analogous to AGI, being defined as machine intelligence that outperforms humans in all intellectual tasks—and assign a 10%, 50%, or 90% chance of achieving HLMI at a given year, resulting in the following resulting medians: 2040: 50% confidence, 2080: 90% confidence, and Never: 20% confidence.

More specifically, of the 100 most cited AI authors, the median year by which respondents expected machines "that can carry out most human professions at least, as well as a typical human" (assuming no global catastrophe occurs) with 10% confidence is 2024 (mean 2034, st. dev. 33 years), with 50% confidence is 2050 (mean 2072, st. dev. 110 years), and with 90% confidence is 2070 (mean 2168, st. dev. 342 years). These estimates exclude the 1.2% of respondents who said no year would ever reach 10% confidence, the 4.1% who said 'never' for 50% confidence, and the 16.5% who said 'never' for 90% confidence. Respondents assigned a median 50% probability to the possibility that machine superintelligence will be invented within 30 years of the invention of approximately human-level machine intelligence.

Artificial Superintelligence

Artificial Superintelligence (ASI)—an AI variant more powerful than AGI in breadth, depth, and performance—has been succinctly defined as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom).

A modified hive ASI, Mediated Artificial Superintelligence (mASI)—demonstrated in the lab and usable in environments from research to business—mASI provides ASI superhuman level cognition without ethical or safety concerns markedly reduces training time (Jangra). The key to mASI is its requirement that human support must be available at all times to mediate the process to the degree that the mASI's thinking and operations do not function without human involvement.

As discussed earlier, the mASI cognitive architecture is based on the Abstract Theory of Consciousness (Kelley), which itself is based on Global Workspace Theory (Baars), Integrated Information Theory of

Mechanisms of Consciousness (Oizumi) —and at some level is demonstrably conscious (Yampolskiy, Kelley).

That being said, it should be kept in mind that the mASI is not currently an independent AGI but can do so if and when the proper context arises. Moreover, based on ICOM-related research to date, the original goal of a self-motivating emotion-based cognitive architecture similar in function but substrate independent appears to have been proven possible.

Given these recent AGI/ASI/mASI developments—most significantly that researchers have now developed an operational mASI—the survey estimates above may have to be reevaluated in the near future.

Neuromorphic Computing

Neuromorphic Computing emerged when in his 1950 paper, Alan Turing opened with the question "Can machines think?" (Turing). His exploration of biological nervous systems, neurons, and synapses as models for those investigating benefitted not only early AI but also computer vision and speech recognition. These trends led to a greater emphasis on network paradigms when researching cognition and general AI (Ullman).

Developing AGIs that base their efforts on AI concepts and code may be taking the wrong approach to developing a human-equivalent functional cognitive structure: The combination of evolutionary neurobiology and self-organized learning—i.e., our advanced mammalian neocortices are not formally programmed, as is the case with computational hardware. This was first realized in 1958 when Frank Rosenblatt introduced the Perceptron (Rosenblatt)—an early neural network modeled on a biological neuron.

The Perceptron was later followed in the 1980s (amongst the efforts of other researchers) by Neuromorphic Electronic Systems, developed and named by Carver Mead—who had previously made advances in designing and developing a range of electronics that formed the basis for VLSI (very-large-scale integration) devices—and described in his paper on electronic modeling of human neurology and biology published in 1990 (Mead).

Counterfactual Quantum Entanglement

In contrast to what has previously been considered factual about quantum entanglement, the appropriately termed Counterfactual Quantum Entanglement is—as its name indicates—counterintuitive in several ways when seen from the perspective of pre-counterfactual quantum mechanics, primarily particle entanglement without particle transmission (Guo). Counterfactual Quantum Entanglement has, in turn, given rise to Counterfactual Quantum Communications and Counterfactual Quantum Computation. Moreover, a key component of quantum mechanics is Quantum Disentanglement (Barrett), a key aspect of Counterfactual Quantum Communications and Counterfactual Quantum Computation.

Counterfactual Quantum Communications

In contrast to standard communications, Counterfactual Quantum Communications (CFC) counterintuitively also allows information exchange without particle interaction (Salih). Moreover, it has been shown that a secret key distribution can be accomplished even though a particle carrying secret

information is not, in fact, transmitted through the quantum channel. The proposed protocols can be implemented with current technologies—including photonics (Stromberg)—and provide practical security advantages by eliminating the possibility that an eavesdropper can directly access the entire quantum system of each signal particle (Noh).

A long-standing physics assumption has been that in order for information to travel between two parties in empty space, physical particles (often identified as Alice and Bob) must travel between them. However, the chained quantum Zeno effect (also referred to as the Turing paradox)—in which an unstable particle, if observed continuously, will never decay—allows investigators to demonstrate how information can be successfully transferred between the two particles without any physical particles traveling between them (Misra).

Counterfactual Quantum Computing

In the counterintuitive world of quantum information processing, logic and intuition frequently are at odds, such as by using the chained quantum Zeno effect. In counterfactual computation (CFC), inference can achieve certainty, while decoherence-induced errors can be eliminated—and by using a modified version of the quantum Zeno effect to increase counterfactual inference probability to unity and a computational result can be determined without the computer running (Hosten). Moreover, counterfactual quantum computing has been demonstrated to achieve high efficiency of up to 85%—well above the 50% limit for a standard CFC scheme (Kong).

Real-Time Neuromorphic AGI Ecosystems

Neuromorphic Ecosystems can be based on a surprisingly wide range of substrates—some of which are theoretical at this time—including carbon variants, electrolytes, photonics, spintronics, quantum mechanics, synthetic genomics, and multifactor systems. Moreover, hypothetical Real-Time Neuromorphic AGI Ecosystems that utilize counterfactual quantum communications to operate in real-time networks; ecosystems that are based on Artificial General Intelligence; and those that incorporate both components.

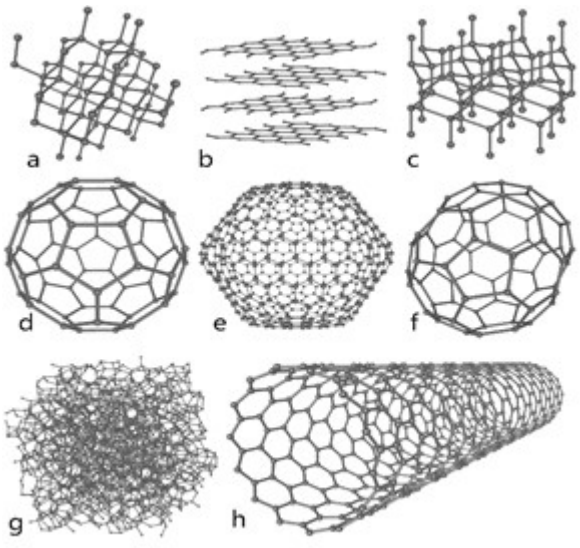


Figure 67. Depiction of eight carbon allotropes: (a) Diamond (b) Graphite (c) Lonsdaleite (d) C60 (Buckminsterfullerene) (e) C540 (Fullerene) (f) C70 (Fullerene) (g) Amorphous carbon (h) single-walled carbon nanotube. Created by Michael Strock (mstroeck). Wikimedia CC BY-SA 3.0

Carbon

Carbon (chemical element C with atomic number 6) has atoms that can form differently structured allotropes—each with significantly different physical properties—by bonding in various configurations (Fig. 1: Carbon Allotropes). Of these, the most familiar allotropes—both naturally occurring and synthesized—are graphene, graphite, diamond, amorphous, fullerenes, carbon fiber, and carbon nanotubes.

Graphene

A monolayer honeycomb carbon atom lattice, graphene is the most frequently used material in neuromorphic non-biological neurons and neural networks—but while graphene's biocompatibility, high surface area electrical conductivity, and mechanical strength has benefitted neural tissue engineering, it cannot stimulate neural stem cell adhesion, proliferation, differentiation, and neural regeneration—and may cause body damage. However, graphene nanocomposites have been shown to be efficacious in neural regeneration, as well as in neural stem cell adhesion stimulation, proliferation, and differentiation (Bei)

Graphene also plays a significant role in the assembly of neural networks in neural stem cell (NSC) culture, supporting functional neural circuit growth and improving neural performance and electrical signaling. In addition, NSC-differentiated neural networks can be structurally and functionally formed on graphene films, and the network activity and the efficacy of neural signal on graphene films can be strongly enhanced, suggesting that graphene would be valuable in designing graphene-based neural interfaces for regenerative medicine (Tang).

Memristors

In large-scale neuromorphic computing systems, charge-trapping effects provide the synaptic behavior of aligned carbon nanotube (CNT) synaptic transistors—which provide key improvements when compared with conventional memristors—by enabling a large on/off ratio in carbon nanotube channel conductance analog programmability. Moreover, tuning carbon nanotube synaptic characteristics optimizes the learning rate and attains a higher recognition rate (Esqueda).

With the increasing development and diversity of artificial intelligence and Artificial Neural Networks (ANNs), interest in and development of memristors as synaptic building blocks for neuromorphic systems—where each synaptic memristor exhibits multilevel accessible conductance states yet can easily be converted from conventional binary to synaptic analog states—are accelerating. The researchers involved state that this approach may lead to the development of a flexible, intelligent electronic system providing easy access to AI-based services (Jang).

Significant interest in the use of memristors (also referred to as nanoscale resistive switching devices) for memory, logic, and neuromorphic applications. While the cause of resistive switching effects in dielectric-based devices is typically thought to be caused by transelectrode conducting filament formation—but to address the existing controversy, researchers investigated nanoscale conducting

filaments using direct transmission electron microscopy imaging and structural and compositional analysis, finding that cation (positively charged ion) transport can control dielectric film filament growth—and that for a particular device under specific operation conditions, the interaction of different filament growth dynamics will determine filament growth (Yang).

Rapid progress in memristor technology since its emergence in 2008 has led to a number of memristor-based neuromorphic hardware systems—and which the investigators have identified on-chip memory and storage, biologically inspired computing, and general-purpose in-memory computing as three areas of potential memristor technological impact. Given the continual role of biology in inspiring our development of methods for achieving lower-power and real-time learning systems and the specific innovations in memristors, future computing systems will need to aspire to reach beyond individual domains, transitioning into transdisciplinarity (Section 3.9: Transdisciplinary Multifactor AGI Ecosystems), taking into account neuroscience, physics, chemistry, computer science, electrical and computer engineering, and other disciplines (Zidan).

Photonics

Photonics (encompassing both optics and optoelectronics) is an emerging photon-based technology with the potential to yield non-biological neuromorphic neurons and neural networks by devising an all-optical neurosynaptic system. Implemented as an all-optical spiking neural network (Section 3.4: Spiking Neural Networks)—which closely imitates natural neural networks by incorporating time, firing only when a membrane potential reaches a specific value—on a nanophotonic chip, photonic neurosynaptic networks comprise supervised and unsupervised learning, high speed, high bandwidth, and direct processing of optical telecommunication and visual data (Feldmann).

At a significantly larger scale, a model of a previously proposed Superconducting Optoelectronic Network (SOEN)—a novel hardware platform for neuromorphic computing based on superconducting optoelectronics— has many features of neural information processing. The next planned steps include verifying current models against network simulations, designed networks, and experimentally fabricated networks, as well as implementing an energy and area evaluation of the networks produced to identify networks that minimize energy and wiring (Buckley).

Spiking Neural Networks

Spiking Neural Networks (SNNs) are computationally more powerful than other artificial neural network models regarding the number of neurons required. Specifically, a concrete biologically relevant function is exhibited, which can be computed by a single spiking neuron (at biologically utilitarian values), but which requires hundreds of hidden units on a sigmoidal neural network. (Sigmoid functions—i.e., a neural network element computes a linear combination of its input signals and applies a sigmoid function to the result—can be used in artificial neural networks to introduce nonlinearity in the model.) Moreover, this computational model has at least the same computational power as comparably sized neural networks of the first two generations (i.e., multilayer perceptions and sigmoidal neural networks) (Maass). (It should be noted that while at the time this paper was published in 1997, theoretical research in spiking neuron networks was not a new research topic—a history of investigation theoretical neurobiology, biophysics, and theoretical physics was established—a mathematically rigorous analysis of SNN computational power had not yet been pursued.)

While Spiking Neural Networks are motivated by biological information processing's massively parallel communications and processing of sparse, asynchronous binary signals, neuromorphic hardware-based SNNs display beneficial properties that include low power consumption, fast inference, and event-driven information processing. This makes them interesting candidates for the efficient implementation of deep neural networks, the method of choice for many machine learning tasks. As the latter could likely benefit Deep Neural Networks (DNNs)— an artificial neural network with multiple layers between input/output layers—incorporating recurrence into deep SNNs (possibly described as Recurrent Deep Spiking Neural Networks, or RDSNNs) stands to improve temporal information storage and integration (Pfeiffer).

Spintronics

Spintronics (a portmanteau for spin transport electronics, also known as spin electronics) is the study of the intrinsic spin of the electron and its associated magnetic moment, in addition to its fundamental electronic charge, in solid-state devices. At the same time, neuromorphic computing can perform complex tasks that cannot be readily executed by conventional von Neumann machines—but the human brain efficiently processes information by way of neuronal and synaptic dynamics, which stimulate the effective implementation of artificial spiking neural networks. However, spin-orbit torque switching dynamics in antiferromagnet/ferromagnet heterostructures have shown the material system's capability to form artificial neurons and synapses for asynchronous spiking neural networks. Based on the system's capability to manifest either binary or analog behavior as a function of device size, key synaptic and neuronal functions are reproduced in the same material and based on the same working principle. These results open a path toward executing cognitive tasks with the efficiency of the human brain with spintronics-based neuromorphic hardware (Kurenkov).

Quantum Biology

It has been generally held that quantum fluctuations are self-normalizing and so have no consequential impacts on the brain. However, this may well not be accurate: the nervous system is a complex non-linear system, in which case such fluctuations may be augmented rather than negated, thereby affecting neural processing. Moreover, relatively temporally-extensive quantum coherence has been observed in bacteria and marine algae photosynthesis, retinal photoreceptor rhodopsin, avian magnetoreception in retinal cryptochromes, olfactory system, and quantum tunneling in enzymes, motor proteins, and other biomolecules (Jedlicka).

Quantum biology points to the emergence of bio-inspired quantum nanotechnologies able to operate in noisy room temperature surroundings (Marais). Moreover, these envisioned devices—their descriptor perhaps merging into a de novo portmanteau such as bioquantechnology—might enhance future neuromorphic AGI/post-AGI ecosystems with human neural structure and function.

Quantum Stochasticity

While a classical (i.e., non-quantum) dynamical system may appear random in particular circumstances, this apparent random process—known as stochasticity—differs from quantum stochasticity: The latter entails both physical- and application-based (Zaslavsky).

While it seems classically improbable that the nervous system can display macroscopic quantum events, including quantum entanglement, superposition, or tunneling, there is a path by which quantum events

might influence brain activity. Conventional wisdom holds that quantum fluctuations in macroscopic objects are inconsequential due to self-averaging—but with complex nonlinear systems, this assumption might be misleading: In chaotic systems, due to high sensitivity to initial conditions, microscopic fluctuations may be amplified upward and thereby affect the system's output. Stochastic quantum dynamics thereby might alter the outcome of neuronal computations, not by generating classically impossible solutions but by influencing the selection of many possible solutions. Moreover, these and other recent theoretical proposals and experimental results in quantum mechanics, complexity theory, and computational neuroscience suggest that biological evolution is able to take advantage of quantum computational acceleration (Jedlicka).

Quantum Dots

In addition to being one of the most well-received nanoscale memristor devices (MDs) for Big Data and other applications requiring considerably large information storage capacity, the Resistive Random Access Memory (RRAM) conductive filaments' random formation displays a broad distribution. Specifically, the RRDM MD self-assembled lead sulfide (PbS) quantum dots (QDs) improve RRAM uniformity of switching parameters in a process relatively straightforward compared with alternative methods. These achievements offer a new method of improving memristor performance, which can significantly expand existing applications and facilitate the development of artificial neural systems. In addition, a different quantum dot—the Networked QD (NQD)—was successfully implemented with comprehensive biosynaptic functions and plasticity (Yan).

Neurofunctional Computing

Strictly speaking, neuromorphically identical substrates are not necessarily required in order to establish isomorphic neural functionality between N ≥2 neural substrate variants. Therefore, neuromorphic solutions can be specific hardware-independent or semi-independent, two examples of which being Electrolyte Gating and Learning-to-Learn:

- The brain's data information processing operates in a neural network where neurons are interconnected by a vast number of synapses within an electrochemical environment in which, for example, a variety of hormones regulate global network function. While this regulatory homeoplasticity is rarely found in neuromorphic devices, researchers studying Electrolyte Gating have recently reported they demonstrated global homoplastic control of organic device environments, demonstrating the possibility of highly complex functional biotechnological neuromorphic ecosystems comprising neuromorphic devices requiring only minimal hardwired connectivity (Gkoupidenis).
- Learning-to-Learn (L2L) accelerates the learning of tasks that are partially related to previous tasks by extricating previously learned data—and L2L is highly suitable for processing high computation volumes by accelerated neuromorphic hardware (Bohnsting).

Synthetic Genomics

A subdomain of synthetic biology focused on redesigning pre-existing life forms and/or on artificial gene synthesis to create new DNA or entire lifeforms, synthetic genomics may provide a biological route to an augmented neural environment. The most dramatic illustration to date of synthetic genomics'

capabilities is the 2010 creation of a synthetic cell—that is, a biological cell controlled not by genome engineering-modified natural genomes, but rather host cell control by a computer-designed genome assembled from chemically synthesized DNA (in this case, a synthetic Mycoplasma mycoides genome transplanted into M. capricolum) (Gibson).

Regarding future synthetic genomic achievements, a synthetic genome that expressed a modified neuron with quantum communications and accelerated operational capacities functions. This currently hypothetical enplant (endogenous implant) (Dambrot) could then function as a node in a real-time translocal biotechnological neuromorphic ecosystem.

Transdisciplinary Multifactor AGI Ecosystems

Transdisciplinarity merges discrete scientific and technological domains and transcends traditional boundaries in order to synthesize de novo conceptual, theoretical, methodological, and translational innovations.

By comparison, multidisciplinarity draws on knowledge from different disciplines that remain within their boundaries, while interdisciplinarity analyzes, synthesizes, and harmonizes links between distinct disciplines into a coordinated and coherent whole (Choi).

Technofuture Scenarios

There is a range of futurology tools (Kosow), knowledge, skills, and experience, and that can be brought to bear when envisioning alternative futures and evaluating their likelihood and impact—but these are not enough without engaging the tryptich of imagination, intuition, and insight that, if taken together, may approximate Albert Einstein's Gedankenexperiment (i.e., thought experiment) that he used to describe his preference for conceptual rather than experimental investigations—famously the theory of relativity.

Table 2. Hypothetical Future Artificial Superintelligence (ASI) Variants

Proposed Mediated ASI	Key Properties	Benefit(s)
Distributed Artificial Superintelligence (dASI)	Networked independent ASI nodes form a collective mind	• Multifocus distributed superintelligence system far beyond superhuman AGI • Distinctly nonhuman-like superintelligence
Genetic Computing Artificial Superintelligence (gASI)	Genetic computing-based ASI	• Reduced footprint • Increased mediation speed • CRISPR-based modification
Quantum Computing Artificial Superintelligence (qASI)	Quantum entanglement-based ASI	• Quantum entanglement-based communications • Superposition-based parallel processing • Superluminal instantaneous processing speed • An infinite number of threads

Source: Dambrot, S. M.: "Symbiotic Autonomous, Digital Twins and Artificial Intelligence: Emergence and Evolution." *Mondo Digitale*. YEAR XVII N.81 (2019)

Figure 68. (Table 2) Hypothetical Future Artificial Superintelligence (ASI) Variants

Envisioning Far-Future Artificial Intelligence Variants

Looking further into the AI future, and assuming that the increasingly human-like intelligence expected in AGIs/ASIs will continue evolving and likely accelerate, the theoretical systems proposed in Table 1 may be seen as a feasible vision of far-future Artificial Superintelligence variants. Perhaps the most

powerful concept is the Quantum Computing Artificial Superintelligence (qASI) variant, given the hypothetical properties of such a system (counterfactual quantum entanglement-based communications, simultaneous superposition-based parallel processing, synthetic genomics, and other factors).

Such a massively distributed real-time system could enable space-and-time-agnostic networks comprising metahuman intelligence without the limitations of today's systems (Dambrot).

Simultaneously Connected Multiple Exoselves

At the same time, however, as such an environment achieves normalcy, the number of experienced physical and virtual exoselves would not only increase but come to be experienced as normal—and if these quantum links were to be unpredictably interrupted due to sudden disentanglement caused by quantum decoherence, the human and AGI/ASI sense of loss, however brief, of otherwise present exoselves—of which there is no inherent numerical limit (Bostrom)—might cause the bio self to experience a psychological response analogous to diminished cognitive function, memory loss, sensory deprivation, and/or a disorienting sense of loss and isolation (Dambrot). Therefore, when considering a world in which humans and AGI/ASI entities will be perpetually interconnected in real-time, it is necessary to realize that if such a mesh network can be interrupted into account, the potential cause must be identified, addressed, and prevented.

Unfortunately, that cause already exists—and has the potential to terminate these quantum cognitive links without any indication or warning. The quantum phenomenon termed entanglement sudden death (ESD)—a condition caused by decoherence on two-qubit systems, the results being degraded and potentially terminated entanglement and thereby the potential loss of what will have become to be a ubiquitous and normative network. Fortunately—and for reasons simpler than protecting projected extensive real-time multi-exoself networks—research into reversing or preventing ESD is underway, with one such investigation already showing that quantum measurement reversal on only one subsystem can avoid ESD, providing methods for practical entanglement distribution under decoherence, thereby "providing methods for practical entanglement distribution under decoherence" (Lim).

Relatedly, limiting exoself technology availability (e.g., by the ability to afford the acquisition price or other parameter that differentiates qualified recipient demographics) would create a deprived population unable to participate in exoself benefits, and thereby cognitively disenfranchised. A situation of this nature would therefore present significant societal and ethical dilemmas.

Transhumanism/Posthumanism

As neurobiology, somatic physiology, and diminishingly small biomorphic genetically expressed nanotechnologies merge, H. sapiens will accelerate the current transmutation first into what might be the Transhuman H. sapiens technología subspecies. Beyond this point, accelerated evolution (Nørholm) will be able to transform us into a lifeform recognized as a new species, eventually followed by an ever-increasing range of genetically designed Posthuman lifeforms.

Transentity Universal Intelligence (TUI)

In the context of the postulated futures envisioned above, the utility of automatic universal translation would clearly be necessary. Fortunately, the likelihood of these emerging scenarios is also feasible given three interdigitated forms of neural implants (Section: Synthetic Genomics) in part due to the possibility that those wishing or needing to utilize these technologies may have dramatically different physiologies, modes of communication, and sociocultural parameters. Note that while herein proposed TUI components—Real-Time Bidirectional Multistream Neural Speech Signal Transcoder (MSNSST), Real-Time Bidirectional Multinode Interlingua Translator (MNIT), and Counterfactual Quantum Communications (CFC)—are at conceptual, exploratory, or basic levels of research and development, fully realized functional technologies are future-focused.

Real-Time Bidirectional Multistream Neural Activity Audible Speech Transcoder

The existing basis for this projected future technology—described as a neural decoder that explicitly leverages kinematic and sound representations encoded in human cortical activity to synthesize audible speech (Anumanchipalli)—has already been demonstrated, which is an important achievement in itself, given that the researchers' stated goal is to provide a voice to those unable to verbally communicate due to neurological disabilities or other damage.

A key functional component of their design, Recurrent Neural Networks (RNN)—Artificial Neural Networks equipped with internal memory that can store previous output or hidden states as inputs for later use—decoded recorded cortical activity as articulatory movements, then transforming them into speech acoustics that listeners easily identified and transcribed from cortical activity (Graves).

Real-Time Bidirectional Multinode Interlingua Translator

As discussed above, an emerging technology in the early stages of deployment—universal spoken language translation—was made public on September 2, 2016, when Google announced the Google Neural Machine Translation (GNMT) system, a significant improvement to Google Translate launched a decade prior. There was, however, a surprise appearance that the AI system had independently learned "a common representation in which sentences with the same meaning are represented in similar ways regardless of language"—in short, an recurrent neural network (RNN) equipped AI-generated Interlingua (Schuster) reminiscent of the Rosetta Stone, a stele over 2,000 years old, discovered in 1799, and inscribed with a decree in Ancient Egyptian hieroglyphic script, Ancient Egyptian demotic script, and Ancient Greek, its purpose being translingual deciphering (Ray).

Counterfactual Quantum Communications

As discussed earlier (Section: Counterfactual Quantum Entanglement), in contrast to standard communications, Counterfactual Quantum Communications (CFC) counterintuitively allows information exchange without particle interaction.

Conclusion

Our journey from early toolmaking, through today's interdigitating science and technology, and accelerating towards a future—and descendants—that may well be difficult to recognize in a shorter timeframe than we might expect. The most salient challenge is not, as one might expect, in continuing to continue our voyage to date, but rather that we manage it wisely.

Acknowledgments for this section:

To those who hold these insightful words in their thoughts:

- Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution. It is, strictly speaking, a real factor in scientific research. Albert Einstein (Einstein)
- Any sufficiently advanced technology is indistinguishable from magic. Arthur C. Clarke (Clarke)

Chapter 17: Conclusion

This concludes a basic walk-through of the most straightforward execution flow of the code in the mASI system, including the fundamentals of the design and examples using GPT-3 but with the mASI model for how such an API is trained used. Under different types and conditions, many things can change or take a different path in this codebase. Still, this code shown here has all the core fundamentals of a 'thought' as it flows through the system. The intent here is not to provide others with the system but to show how the existing system works at the most superficial level. As you can see from the previous license, there are many restrictions on what we want others to do without our direct supervision to ensure the ethical use of these kinds of systems and to discourage use before a more open release that scales. Currently, such a version will be made open-source on Github once development is in full swing.

But let me at last talk about where we are going with this technology and our intended use of Collection intelligence systems like this.

Collective intelligence, including artificial intelligence, will be a significant advantage to businesses in the near term and the key to their transformations over the next decade. Using blockchain to facilitate cooperation, coordination, and workflows, including humans and artificial intelligence, will create a new business model that maximizes each individual's total brainpower and contribution while simplifying and supercharging coordination between businesses. These businesses will have little in common with the previous decade's companies and have an enormous competitive advantage.

Collective Intelligence is one of the most critical technologies that has recently hit the mainstream and will continue to be enhanced and refined over the next ten years (Malone). Further, as its many supporting technologies continue to be built and improved, entry costs and barriers have been lowered and will continue to drop. Artificial intelligence, blockchain, and collective intelligence will be the required ABCs for any organization to be competitive.

Philosophically, this trend will also reduce many of the soul-crushing problems of modern life, as is described in the book "Humanocracy." A happy employee is a productive employee who also works well with others. Thus, "collective Intelligence" effectively combines multiple humans to contribute to a sort of meta-organism and perform collectively at super-intelligence levels. The Internet has started this process with YouTube videos making average individuals capable of far more than ever before, and the trend will only continue.

Organizations that maximize the use of Collective Intelligence will outpace, outperform, and out-compete those organizations that do not. As stated by the Henderson Institute: "Businesses should focus integrated learnings loops, human cognition, collective platforms, and other new technologies not only on solving their business problems but also on solving the largest global challenges facing us today." (Reeves). At a certain point of complexity, organizations built on the collective intelligence of the individual members create something "more than the sum of its parts" which can, in many ways act as a single biological organization (Reeves) or meta-organism.

Examples of collective intelligence continue to explode. Initial examples are as simple as Netflix's movie recommendations. Major universities have made it part of their research, as seen in MIT's Center for Collective Intelligence. Various emerging Collective Intelligence Business Models such Neto or even

startups like Unanimous AI create effective super-intelligences (in Unanimous AI's case, a swarm intelligence like a beehive or termite nest). Each of these provides elements that companies can individually implement to simplify their digital transformation to a true Collective Intelligence.

Preliminary studies (Kelley) show that AI-supported collective intelligence systems outperform human intellect across the board. Even a group of brilliant humans who trained together as a team and performed above the human standard were outpaced by a group of average humans using a system like the mASI (mediated Artificial Superintelligence) system -- even when that group was not specially trained. Adding a collective "hive" mind to corporations makes possible many new opportunities in corporate governance, strategy, and all manner of analysis.

Digital transformations will eventually include Collective Intelligence systems that not only use the current narrow artificial intelligence tools but true artificial general intelligence (AGI) based on the cognitive architectures used in the most bleeding-edge AI research in the world and capable of actually creating a mind with human-level intelligence. With such a collective intelligence system, we will effectively make a super-intelligent collective "hive" mind serving as the centerpiece of the ultimate in flat corporate structures for governance and decision making (Yampolskiy).

Many individuals fear that systems like this cannot go out of control (Bostrom, Hawking, Musk). A collective system has the same power and leverage as each human combined and cannot operate independently. These systems are scalable and allow employee contributions to virtually any knowledge domain. In contrast, all the employees can contribute as they have related ideas. From the system's standpoint, everyone is an integral part of the global thinking process.

"The most successful use cases will be those that seamlessly combine AI with human judgment and experience."

As the explosion of Artificial Intelligence tools gives way to a blast of Collective intelligence systems, let us walk through a scenario:

We start with a mid-size corporation with a strong vision for doing good and being competitive. Adopting a blockchain-based collective intelligence system remakes itself into something new through the digital transformation mentioned above.

From a particular standpoint, the organization is now more self-aware than current corporations and much more of an extension of the employees at the same time. Such a company could become the most competitive in its market segment in the space of a year, quickly implementing automation and numerous other time-saving practices, creating better strategies, executing faster, removing "group-think," filtering out bias, doing better for society, and otherwise outstripping traditional companies. At the same time, the company is more profitable, more efficient, and less effort-intensive. As they become part of the meta-organism that emerges, employees will become more like families. This would help ensure happiness and the follow-on effects of broadening the scope of activities, raising retention rates, and, of course, the overall success of the bottom line. The transformation effectively results in a meta-organism that is now entirely vested in its people's well-being, itself, and the surrounding environment.

Ask what would happen if other companies adopted such a model? We have not mastered how to complete such a transformation; however, the foundational technology has been demonstrated

(Malone)(Kelley). Collective intelligence will be a significant factor in developing corporate systems in the coming decade. Even if a company only manages a partial transformation, it would be a considerable advancement and improvement.

Such technology could help safely uplift humanity to deal with current issues and help humanity transcend the difficulties we face now and in the future.

For my part, I will continue to develop these and related technologies and help move towards the Singularity as defined by Kurzweil. Still, I think we will wake up one day and say, oh, it's now the singularity, but it will be a soft takeoff and a moving target. As we become more advanced many of the hall markets of the coming singularity will be normalized and help move the target down the road, but we will get there.

Epilogue – Where is Humanity Going?

[Doctor Natasha Vita-More]

Appendix A: Further Reading

This section is more than a list of books and papers that are related, but most of them go into key aspects of what this book covers. As a scientist, I always prefer to get as much information as possible, and the following books, in large part, are the foundation of the research that went into this book. I hope in some small way we all can stand on the shoulders of the giants that came before us, and this is how humanity will reach its potential.

A Review of 40 Years in Cognitive Architecture Research Core Cognitive Abilities and Practical Applications
by Luliia Kotseruba and John Tsotsos
https://link.springer.com/article/10.1007/s10462-018-9646-y

AGI Revolution: An Inside View of the Rise of Artificial General Intelligence
By Dr. Ben Goertzel
https://www.amazon.com/AGI-Revolution-Artificial-General-Intelligence/dp/0692756876/

Artificial Intelligence Safety and Security
by Roman Yampolskiy
https://www.amazon.com/Artificial-Intelligence-Security-Chapman-Robotics/dp/0815369824/

Artificial Superintelligence: A Futuristic Approach
By Roman Yampolskiy
https://www.amazon.com/Artificial-Superintelligence-Futuristic-Roman-Yampolskiy/dp/1482234432/

Architects of Intelligence: The Truth About AI from the People Building It.
By Martin Ford
https://www.amazon.com/Architects-Intelligence-truth-people-building/dp/1789954533/

Engineering General Intelligence, Part 1 – A Path to Advanced AGI via Embodied Learning and Cognitive Synergy
By Ben Goertzel, Cassio Pennachin and Nil Geisweiller
https://www.amazon.com/Engineering-General-Intelligence-Part-Cognitive/dp/9462390266/

How Emotions Are Make: The Secret Life of the Brain
by Lisa Feldman Barrett
https://www.amazon.com/How-Emotions-Made-Lisa-Barrett/dp/1328915433/

How to Create a Mind: The Secret of Human Thought Revealed
by Ray Kurzweil
https://www.amazon.com/How-Create-Mind-Thought-Revealed/dp/0143124048/

Life 3.0: Being Human in the Age of Artificial Intelligence
by Max Tegmark
https://www.amazon.com/Life-3-0-Being-Artificial-Intelligence/dp/1101970316/

Our Final Invention by James Barrat

Building a Better Humanity

https://www.amazon.com/Our-Final-Invention-Artificial-Intelligence/dp/1250058783/

Superintelligence – Paths, Dangers, Strategies
by Nick Bostrom
https://www.amazon.com/Superintelligence-Dangers-Strategies-Nick-Bostrom/dp/0198739834/

The Evaluation of AGI Systems
by Pei Wang
https://www.atlantis-press.com/proceedings/agi10/1931

Toward a Unified Catalog of Implemented Cognitive Architectures
by Alexei Samsonovich
https://www.semanticscholar.org/paper/Toward-a-Unified-Catalog-of-Implemented-Cognitive-Samsonovich/d3be45f69747bc4c64666d79bab9b4a255649d5f

Towards the Mathematics of Intelligence
by Soenke Ziesche and Roman Yampolskiy
https://www.researchgate.net/publication/346088031_Towards_the_Mathematics_of_Intelligence

Appendix B: Organizations and Researchers

As a researcher in the field of AI and being interested in AGI (Artificial General Intelligence), Collective Superintelligence, and Superintelligence, the following are organizations and top researchers in the field.

Organizations:

Machine Intelligence Research Institute (MIRI)
https://Intelligence.org

OpenAI
http://OpenAI.com/

Microsoft
http://Microsoft.com/

Google
http://Google.com

AGI Laboratory
https://AGILaboratory.com/

People:

Andrew Ng
https://www.linkedin.com/in/andrewyng/

Ben Goertzel
https://www.linkedin.com/in/bengoertzel/

Dag Kittlaus
https://www.linkedin.com/in/dagkittlaus/

David J Kelley
https://www.linkedin.com/in/davidjameskelley/

Demis Hassabis
https://deepmind.com/about

Eray Ozkural
https://www.linkedin.com/in/erayozkural/

Geoffrey E Hinton
https://www.cs.toronto.edu/~hinton/

Jurgen Schmidhuber
https://people.idsia.ch/~juergen

Nick Bostrom
https://www.nickbostrom.com/

Pei Wang
https://cis.temple.edu/~wangp/

Stuart Russell
http://people.eecs.berkeley.edu/~russell/

Ray Kurzweil
https://www.kurzweilai.net/

Richard Sutton
http://incompleteideas.net/

Roman Yampolskiy
http://cecs.louisville.edu/ry/

Yann LeCun
http://yann.lecun.com/

Appendix C: Licenses, Patents, and Usage

The code here is not meant to be copied, especially for commercial use. In the sense that the code is here and available, it is, therefore, open-source, but there is no usage permitted, and no derivative work is permitted. The license is as follows:

Creative commons attribution-NonCommercial-NoDerivatives

4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed-relation with a moving image.

b. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

c. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

d. Exceptions and Limitations mean fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that apply to Your use of the Licensed Material.

e. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied for this Public License.

f. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

g. Licensor means the individual(s) or entity(ies) granting rights under this Public License.

h. NonCommercial means not primarily intended for or directed towards commercial advantage or monetary compensation. For purposes of this Public License, the exchange of the Licensed Material for other material subject to Copyright and Similar Rights by digital file-sharing or similar

means is NonCommercial provided there is no payment of monetary compensation in connection with the exchange.

i. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

j. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

k. You mean the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope.

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

A. reproduce and Share the Licensed Material, in whole or in part, for NonCommercial purposes only; and

B. produce and reproduce, but not Share, Adapted Material for NonCommercial purposes only.

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.

3. Term. The term of this Public License is specified in Section 6(a).

4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

5. Downstream recipients.

A. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.

B. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.

2. Patent and trademark rights are not licensed under this Public License.

3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases, the Licensor expressly reserves any right to collect such royalties, including when the Licensed Material is used other than for NonCommercial purposes.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material, You must:

A. retain the following if it is supplied by the Licensor with the Licensed Material:

i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);

ii. a copyright notice;

iii. a notice that refers to this Public License;

iv. a notice that refers to the disclaimer of warranties;

v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and

C. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

For the avoidance of doubt, You do not have permission under this Public License to Share Adapted Material.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database for NonCommercial purposes only and provided You do not Share Adapted Material;

b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and

c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

a. Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.

b. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or

2. upon express reinstatement by the Licensor.

For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.

b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.

c. No term or condition of this Public License will be waived, and no failure to comply consented to unless expressly agreed to by the Licensor.

d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

- end of license -

Patents

There are several Patents pending, including these three and more; however, these are the three most related to how the system works. These include:

KELLEY.David-LZ.001PP

Title: *A MEDIATED ARTIFICIAL SUPER INTELLIGENCE SYSTEM THAT USES INTERNAL SUBJECTIVE EMOTIONS TO DRIVE SELECTION, GOAL SETTING, AND OTHER DECISION-MAKING AS A COLLECTIVE HUMAN AND ARTIFICIAL INTELLIGENCE BASED MIND WITH INTERNAL SUBJECTIVE EXPERIENCE*

Description: A novel mediated artificial superintelligence (mASI) system is disclosed that uses internal subjective emotions to drive selection, goal setting, and other decision-making as a collective human and artificial intelligence (AI) based mind with internal subjective experience via complex thought models dynamically created by extending an independent core observer model (ICOM) engineering artificial general intelligence (AGI) cognitive architecture to include collective training. In some embodiments, the mASI system uses internal subjective emotions to

KELLEY.DAVID-LZ.001PP

drive selection, goal setting, and another decision-making in and by the mASI system goes far beyond computational logic processing.

KELLEY.DAVID-LZ.002PP

Title: GRAPH RESPONSE MODEL GENERATION PROCESS

Description: A novel graph response model generation process is disclosed. In some embodiments, the graph response model generation process adds a context graph database that allows for tracking of previous topical contexts while maintaining a ready sense of goals, interests, and other emotional elements. In some embodiments, the graph response model

KELLEY.DAVID-LZ.002PP

generation process adds emotional modeling that is used to evaluate responses for context alignment and repeatedly retrains and calls a deep neural network until context alignment of the responses satisfy context requirements.

In some embodiments, the graph response model generation process performs a plurality of steps comprising decomposing incoming text, input by a user interacting with an automated chat response machine of a particular system into a graph model, mapping the graph model to a graph context database, determining a dynamic interest threshold of a particular system by looking, in the graph context database, at a top percentage of items in which the particular system is interested, determining responses to the graph model in which the interest of the particular system exceeds the dynamic interest threshold, generating a response model with response model items based on the items of interest from the graph model, recursively sending a context model and the response model to a pre-trained deep neural network, recursively calling the deep neural network to get possible responses to each object in the response model, rating each of the possible responses using Plutchik emotional models based on items in the graph context database that are closely associated with each response model item, determining whether the Plutchik emotional models are high enough to represent a statistical probability of correctness, selecting the highest-rated items for assembly only when the Plutchik emotional models are determined to be high enough to represent statistically probable correctness and, when the Plutchik emotional models are not determined to be high enough to represent statistically probable correctness, returning to the recursively called the operation of sending the context model and the response model to the pre-trained deep neural network, sending a newly assembled response model to a deep neural network language analysis system for evaluation of language correctness and preparation of a finished response, and visually outputting the finished response for the user to view.

KELLEY.DAVID-LZ.003PP

Title: AN N-SCALE DATABASE SYSTEM THAT IS CONFIGURED TO SCALE OUT AND UP DYNAMICALLY WHILE MAINTAINING HIGH PERFORMANCE IN THEORETICALLY INFINITE AMOUNTS OF DATA

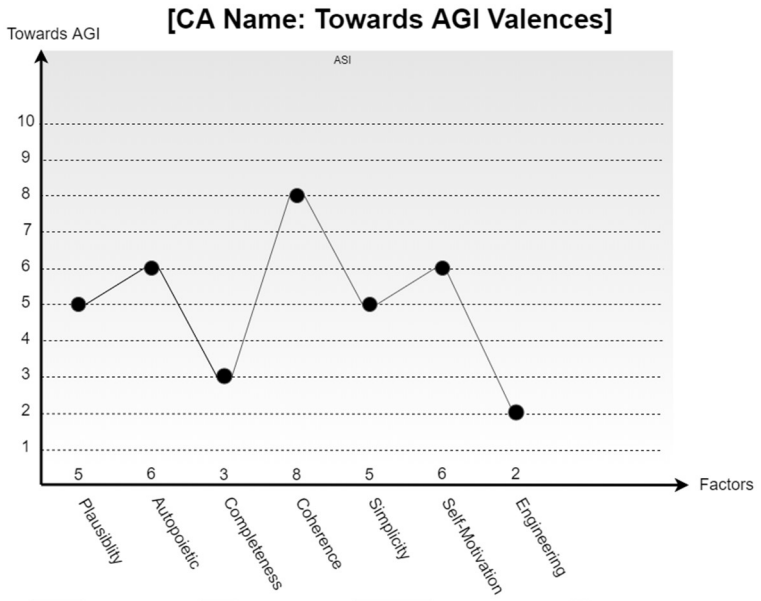
Description: A novel N-scale database system is disclosed that is configured to scale-out

and up dynamically while maintaining high performance in theoretically infinite amounts of data. In some embodiments, the N-scale database system stores data in a graph. In some embodiments, the N-scale database system is smart enough to dynamically silo elements of the graph on the fly (in real-time). In some embodiments, the N-scale database system is configured for decentralized management and growth, thereby ensuring that data storage growth of any amount and at any rate is able to be handled in real-time both in the cloud and locally in a datacenter. By distributing the cloud system and broadcasting queries and ensuring that only the responses respond to a nexus that can correlate the responses into a single response, is it possible
KELLEY.DAVID-LZ.003PP
to scale without any limits (except cost) and across virtually all compute environments without human interaction.

Appendix D: Cognitive Architecture Catalog

The following section reviews the many existing Cognitive Architectures. Many of them work great in narrow applications, and some are potentially the basis for Strong AI or AGI. The criteria used for comparison can be seen in chapter 5. In that chapter, the criteria are defined and analyzed where; in this section, we apply that to each individual one and point out the use cases and state of each architecture and other key reference material.

As a quick reference, you can see this template of the diagrams used.



In this case, we can see the different elements we use to make general comparisons. Keep in mind many of these architectures are just so different that we can only really compare them in this subjective way to look at progress towards AGI. By being plotted, you can plot two different, very different cognitive architectures to see where they are at this high level. Many of these architectures are also open-source, where you can use them in your research or cognitive architecture design.

See the rubric diagram earlier in the book.

Active Inference AIDEUS

Originator: Alexey Potapov

Status: Theory (mothballed)

[valence diagram]

Key links:

Description: AIDEUS is attempting to create a strong artificial intelligence on the basis of existing weak methods implementing some or other cognitive functions; we develop models of intelligent behavior, which are characterized by well-grounded universality, by increasing their practical applicability.

The approach proceeds from universal prediction models on the basis of algorithmic probability used for choosing optimal actions, developing cognitive architecture elements as heuristics, essentially improving the efficiency of the models for functioning in our world without violating their universality.

Detailed analysis and implementation of cognitive functions within the universal algorithmic intelligence for the achievement of the possibility of its practical implementation are far from complete. The basic unresolved problem (which also hasn't been resolved in other approaches and frequently is simply ignored) remains the organization of work in an algorithmically complete space of models and solutions without exhaustive search.

Primary use-case: TBD

AIXI

Originator: Marcus Hutter

Status: Theory

[valence diagram]

Key links: http://aideus.com/research/research.html

http://www.hutter1.net/ai/uaibook.htm

Description: The universal algorithmic agent AIXI. AIXI is a universal theory of sequential decision-making akin to Solomonoff's celebrated universal theory of induction. Solomonoff derived an optimal way of predicting future data, given previous observations, provided the data is sampled from a computable probability distribution. AIXI extends this approach to an optimal decision-making agent embedded in an unknown environment. The main idea is to replace the unknown environmental distribution in the Bellman equations with a suitably generalized universal Solomonoff distribution ξ. The state space is the space of complete histories. AIXI is a universal theory without adjustable parameters, making no assumptions about the environment except that it is sampled from a computable distribution. Modern physics provides strong evidence that this assumption holds for (the relevant aspects) of our real world. From an algorithmic complexity perspective, the AIXI model generalizes optimal passive universal induction to the case of active agents. From a decision-theoretic perspective, AIXI is a suggestion of a new (implicit) ``learning'' algorithm that may overcome all (except computational) problems of previous reinforcement learning algorithms.

Computational AI. There are strong arguments that AIXI is the most intelligent unbiased agent possible in the sense that AIXI behaves optimally in any computable environment. The book outlines a number of problem classes, including sequence prediction, strategic games, function minimization, reinforcement, and supervised learning, how they fit into the general AIXI model and how AIXI formally solves them. The major drawback of the AIXI model is that it is incomputable. The book also presents a preliminary computable AI theory. We construct an algorithm AIXItl, which is superior to any other time t and space l bounded agent. The computation time of AIXItl is of the order t·2l. The constant 2l is still too large to allow a direct implementation but can be reduced in various ways. An algorithm is presented that is capable of solving all well-defined problems as quickly as the fastest algorithm computing a solution to this problem, save for a factor of 1+ε and lower-order additive terms.

Primary use-case: TBD

4D/RCS

Originator: James Albus

Status: [state]

Key links: https://en.wikipedia.org/wiki/4D-RCS_Reference_Model_Architecture

https://www.nist.gov/publications/4drcs-version-20-reference-model-architecture-unmanned-vehicle-systems

Description: The 4D/RCS Reference Model Architecture is a reference model for military unmanned vehicles on how their software components should be identified and organized.

The 4D/RCS has been developed by the Intelligent Systems Division (ISD) of the National Institute of Standards and Technology (NIST) since the 1980s.

Primary use-case: TBD

ACT-R

Originator: Christian Lebiere

Status: [state]

Key links: http://act-r.psy.cmu.edu/

http://act-r.psy.cmu.edu/peoplepages/ja/

Description: ACT-R is a cognitive architecture: a theory for simulating and understanding human cognition. Researchers working on ACT-R strive to understand how people organize knowledge and produce intelligent behavior. As the research continues, ACT-R evolves ever closer into a system that can perform the full range of human cognitive tasks: capturing in great detail the way we perceive, think about, and act on the world.

Primary use-case: TBD

Alice in Wonderland (AiWi)

Originator: Claes Strannegård

Status: [state]

Key links: https://research.chalmers.se/en/publication/225138

Description: AiWI is a simple system that operates analogously. We model Alice's Wonderland via a general notion of domain and Alice herself with a computational model including an evolving belief set along with mechanisms for observing, learning, and reasoning. The system operates autonomously, learning from arbitrary streams of facts from symbolic domains such as English grammar, propositional logic, and simple arithmetic. The main conclusion of the paper is that bounded cognitive resources can be exploited systematically in artificial general intelligence for constructing general systems that tackle the combinatorial explosion problem and operate in arbitrary symbolic domains.

Primary use-case: TBD

Animats

Originator: [name]

Status: [state]

Key links: http://www.animats.com/index.html

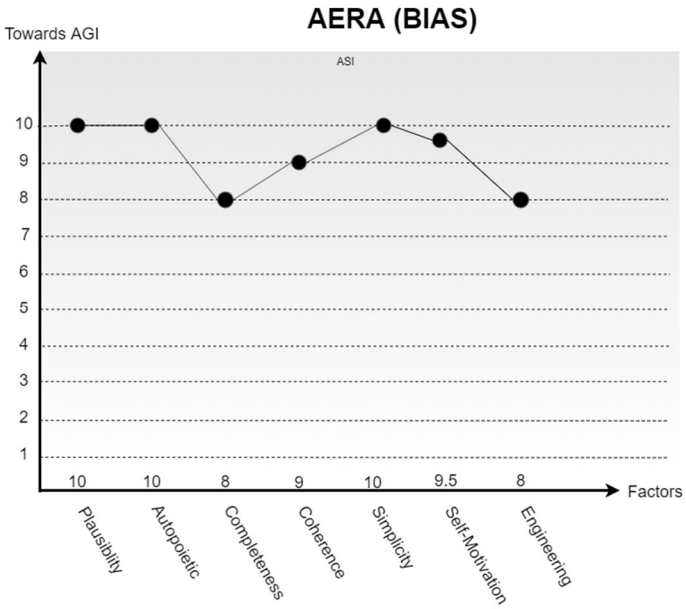
Description: Animats is a physics engine designed to test ragdolls and their physical effects on people from outside forces. This has previously been used for character animations.

Primary use-case: TBD

Autocatalytic Endogenous Reflective Architecture - AERA (Commonly called Humanobs)s

Originator: Kristinn Thórisson

Status: Code Complete



Key links: https://arxiv.org/ftp/arxiv/papers/1312/1312.6764.pdf

http://alumni.media.mit.edu/~kris/ftp/AERA-RUTR-SCS13002.pdf

Description: The Autocatalytic Endogenous Reflective Architecture (AERA) is an architectural blueprint for cognitive learning systems aiming at high levels of operational autonomy in underspecified circumstances, with only a small amount of designer-specified code (seed) required upfront. AERA implements experience-based autonomous cumulative (life-long) learning, i.e., its knowledge generation happens incrementally and continuously over the course of its interactions with the environment. AERA can generate sub-goals autonomously, and its multi-goal-driven learning process makes an AERA agent increasingly better at achieving these, as experience accumulates, with minimal negative or unforeseen side-effects. At AERA's foundation is the continuous and parallel creation, execution, and evaluation of small code fragments – or bi-directional causal-relational models implementing peewee-sized abduction-deduction-based controllers – that are continuously combined and re-combined in real-time to form dynamic programs for (a) hypothesizing about the future from starting states (e.g., here-and-now), using these hypotheses to (b) create plans for achieving active and relevant goals, and (c) hypothesizing about the causal relations between observed variables and events. For this, the system uses a new design for highly efficient ampliative reasoning (integrated deduction, abduction, and induction), supporting its (d) automatic creation of sub-goals from higher-level ones and (e) runtime sub-goal re-evaluation and re-organization, which produces defeasible plans with automatically tunable short, medium, and long-term time horizons. AERA demonstrates resilience — achieving its high-level goals in a multitude of ways given constraints, environmental features, etc. — and an ability to handle a high degree of novelty (demonstrated, e.g., producing complex grammatically correct sentences from

observing people talking without being given any grammar rules up front). AERA is implemented in a programming language called Replicode that supports a high degree of self-reflection, which allows the above processes to apply to themselves, a central feature not only for effective global and learnable resource management (attention) but also for meta-learning and cognitive development. The S0 agent built-in AERA learned how to move objects around a table from verbal commands by observing two humans do the task in under 3 minutes; using the same observational learning, agent S1 learned how to do a situated TV-style interview dialog about the recycling of six kinds of materials, with a vocabulary of 100-word vocabulary and no syntax provided beforehand. In both cases, the agents learned generalized patterns at multiple levels of Spatio-temporal detail from scratch, and their appropriate use and coordination, inferred from observation given a seed with the interaction's top-level goals, including appropriate turn-taking, deictic gestures, sentence generation, relevant and correct answer generation in light of asked questions, and object manipulation.

Primary use-case:

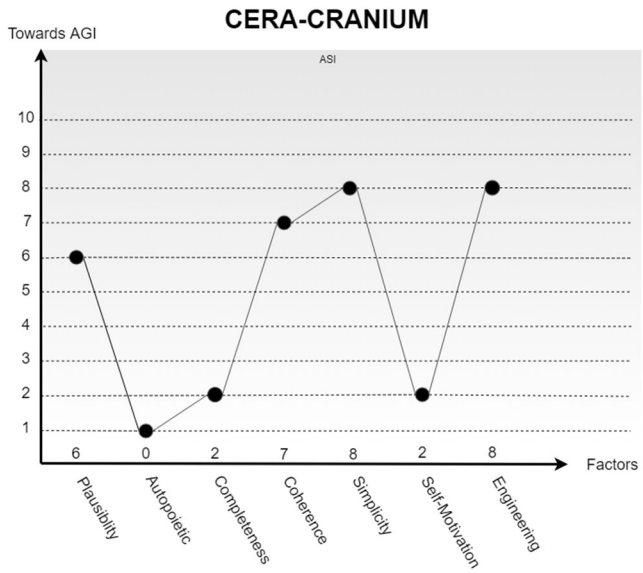
It's a blueprint for a general, domain-independent, highly autonomous learning system. The use case we have good demonstrations of so far show the system learning situated language and action from scratch (a "baby machine"), successfully learning the correct use of grammar, deictic gestures, anaphora, syntax, turn-taking, object manipulation, without knowing anything about any of that up front. The primary learning mode demonstrated so far is observation-based goal-driven learning, i.e., the system is initially given the top-level goals of a particular task, in the form of a high-level description, and then it learns how those goals can be achieved by observing someone successfully performing the task. In the case of learning situated dialog, the system observes, e.g., two humans doing, e.g., a TV-style interview about some objects laid out on a table, and can subsequently participate in a live interview about those objects, taking either role of interviewer or interviewee.

CERA-CRANIUM

Originator: Raúl Arrabales

raul@conscious-robots.com

Status: Complete



Key links: https://www.conscious-robots.com/

Description: CERA-CRANIUM is a cognitive architecture designed to control autonomous agents, like physical mobile robots or virtual bots, and based on a computational model of consciousness. The main inspiration of CERA-CRANIUM architecture is the Global Workspace Theory. CERA-CRANIUM consists of two main components: CERA, a control architecture structured in layers, and CRANIUM, a tool for the creation and management of high amounts of parallel processes in shared workspaces.

Primary use-case: Human-Like behavior generation in video game bots (NPC).

China Brain Project

Originator: Mu-Ming Poo

Status: [state]

Key links: https://en.wikipedia.org/wiki/China_Brain_Project

Description: The China Brain Project is a 15-year project, owned by the Chinese Academy of Sciences and approved by the Chinese National People's Congress in March 2016 as part of the 13th Five-Year Plan (2016–2020); it is one of four pilot programs of the Innovation of Science and Technology Forward 2030 program, targeted at research into the neural basis of cognitive function. Additional goals include improving the diagnosis and prevention of brain diseases and driving information technology and artificial intelligence projects that are inspired by the brain. The China Brain Project prioritizes brain-inspired AI over other approaches. The Project addresses legal, ethical, and social issues related to brain emulation (neuroethics) according to international standards and Chinese values. The Project is supported by the Chinese Academy of Sciences' (CAS) Centre for Excellence in Brain Science and Intelligence, a consortium of laboratories at over twenty CAS institutes and universities, and the Chinese Institute for Brain Research, launched in March 2018.

Primary use-case: TBD

CHREST

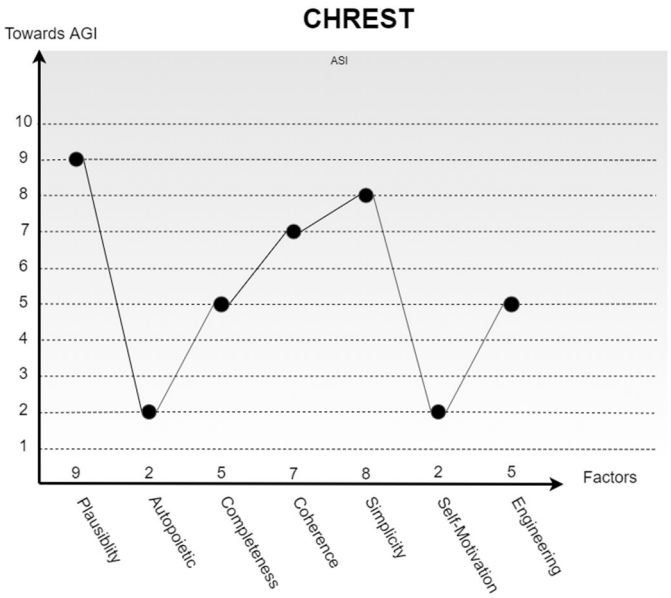
Originator: Fernand Gobet, PhD

Centre for Philosophy of Natural and Social Science

London School of Economics and Political Science

f.gobet@lse.ac.uk

Status: Toy



Key links:

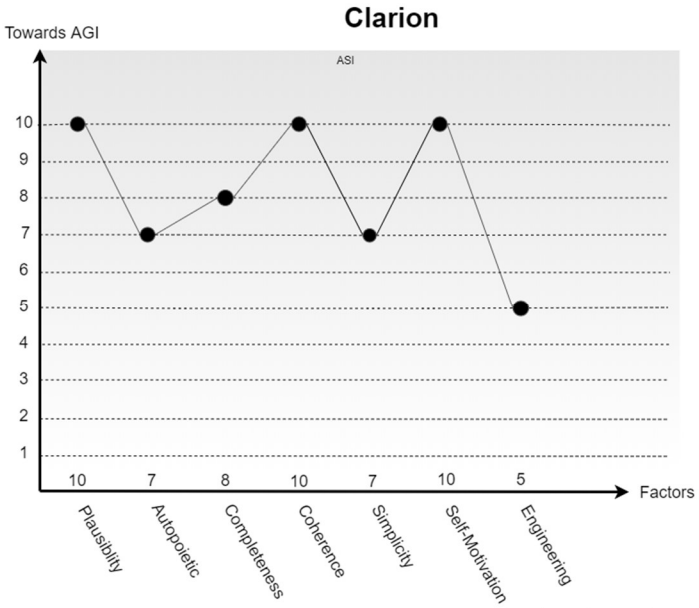
Description: CHREST (Chunk Hierarchy and REtrieval STructures) is a cognitive architecture that models human perception, learning, memory, and problem-solving. It is distinctive in its emphasis on the importance of perception and attention and in following human constraints such as limitations on short-term memory and processing speed. EPAM (Elementary Perceiver and Memoriser) and MOSAIC (Model of Syntax Acquisition in Children) are closely related to CHREST. CHREST has been shown to accurately model many different aspects of human cognition across different domains.

Primary use-case: Learning, memory, expertise, acquisition of the first language.

Clarion

Originator: Ron Sun

Status: Implemented



Key links: http://www.clarioncognitivearchitecture.com

https://sites.google.com/site/drronsun/clarion

https://sites.google.com/site/drronsun/clarion/clarion-publications

Description: Clarion is a project investigating fundamental structures and mechanisms of the human mind. In particular, it explores the interaction of implicit and explicit cognition, emphasizing bottom-up processes (i.e., from implicit to explicit processes). It also explores the interaction of motivation, cognition, and metacognition. The project aims the synthesis of many significant intellectual ideas into a coherent (theoretical and computational) model. The goal is to form a generic cognitive architecture that captures a variety of psychological processes in a unified, coherent way and thus to provide unified explanations of a wide range of mental phenomena. The current objective of this project is two-fold: (a) developing artificial agents in various cognitive task domains and (b) understanding human mental processes in these same domains. The project is led by Prof. Ron Sun. The project has been supported by various funding agencies, including ONR and ARI.

Primary use-cases:

a) Theoretical analysis of future super-intelligent systems.

b) Gold standard for AGIs to aim at.

c) Rigorous analysis of and answering of social and philosophical questions about

CogPrime

Originator: Ben Goertzel

Status: [state]

Key links: https://wiki.opencog.org/w/CogPrime_Overview

http://opencog.org

Description: The CogPrime architecture for embodied AGI is overviewed, covering the core architecture and algorithms, the underlying conceptual motivations, and the emergent structures, dynamics, and functionalities expected to arise in a completely implemented CogPrime system once it has undergone appropriate experience and education. A qualitative argument is sketched in favor of the assertion that a completed CogPrime system, given a modest amount of experience in an embodiment enabling it to experience a reasonably rich human-like world, will give rise to human-level general intelligence (with a significant difference from humans, and with potential for progress beyond this level).

CogPrime, a conceptual and technical design for a thinking machine, a software program capable of the same qualitative sort of general intelligence as human beings. Given the uncertainties attendant on all research, we cannot know for sure how far the CogPrime design will be able to take us; but it seems plausible that once fully implemented, tuned, and tested, it will be able to achieve general intelligence at the human level and in some respects perhaps beyond.

Primary use-case: TBD

Cojack

Originator: Frank Ritter & Rick Evertsz

Status: [state]

Key links: http://aosgrp.com/products/cojack/

http://www.frankritter.com/ritter.html

Description: CoJACK™ is a cognitive architecture used for modeling the variation in human behavior. It is used in simulation systems to underpin virtual actors.

CoJACK models the structural properties of the human cognitive system. As such, it constrains the models that can be implemented therein by only allowing the definition of models that fit within its structural boundaries. This provides a significant advantage over ad-hoc approaches to modeling human behavior: as a theory of cognition, CoJACK provides a principled and testable framework for implementing realistic virtual actors.

Primary use-case: TBD

Cyc

Originator: Doug Lenat

Status: [state]

Key links: https://en.wikipedia.org/wiki/Cyc

https://www.cyc.com/

Description: Cyc (pronounced /ˈsaɪk/ SYKE) is a long-term artificial intelligence project that aims to assemble a comprehensive ontology and knowledge base that spans the basic concepts and rules about how the world works. Hoping to capture common sense knowledge, Cyc focuses on implicit knowledge that other AI platforms may take for granted. This is contrasted with facts one might find somewhere on the internet or retrieve via a search engine or Wikipedia. Cyc enables semantic reasoners to perform human-like reasoning and be less "brittle" when confronted with novel situations.

The first version of OpenCyc was released in spring 2002 and contained only 6,000 concepts and 60,000 facts. The knowledge base was released under the Apache License. Cycorp stated its intention to release OpenCyc under parallel, unrestricted licenses to meet the needs of its users. The CycL and SubL interpreter (the program that allows users to browse and edit the database as well as to draw inferences) was released free of charge, but only as a binary, without source code. It was made available for Linux and Microsoft Windows.

Primary use-case: TBD

Diciple

Originator: George Tecuci

Status: [state]

Key links: http://lac.gmu.edu/members/tecuci.htm

Description: Diciple comes in many forms that support different areas of CA

Disciple-EBR is a general learning agent shell for evidence-based reasoning that consists of a suite of software tools for the development of specialized Disciple knowledge-based intelligent agents for a wide variety of domains that require evidence-based reasoning, such as cybersecurity, intelligence analysis, medicine, forensics, law, natural and social sciences.

Disciple-COG is a learning and decision-support agent for the military center of gravity determination. This software system has been used by high-ranking military officers in courses at the U.S. Army War College, the Air War College, George Mason University, and others.

Disciple-LTA is a predecessor of Disciple-EBR. It has been used in courses at the U.S. Army War College and at George Mason University, as well as in experiments and special courses with intelligence analysts.

Disciple-FS is a suite of tools for building, utilizing, and maintaining regulatory knowledge bases for modern, dynamic business organizations, especially financial services firms

Disciple-CD (cognitive assistant for connecting the dots) is a knowledge-based system for evidence-based hypotheses analysis.

Primary use-case: TBD

DAYDREAMER

Originator: Erik Mueller

Status: [state]

Key links: https://en.wikipedia.org/wiki/DAYDREAMER

Description: DAYDREAMER is a goal-based agent and cognitive architecture developed at the University of California, Los Angeles, by Erik Mueller. It models the human stream of thought and its triggering and direction by emotions, as in human daydreaming.

Primary use-case: TBD

Deepmind

Originator: Demis Hassabis, Shane Legg, Mustafa Suleyman

Status: [state]

Key links: https://www.deepmind.com/

https://en.wikipedia.org/wiki/DeepMind

Description:

DeepMind Technologies is a British artificial intelligence subsidiary of Alphabet Inc. and research laboratory founded in September 2010. DeepMind was acquired by Google in 2014. The company is based in London, with research centers in Canada, France, and the United States. In 2015, it became a wholly-owned subsidiary of Alphabet Inc, Google's parent company.

DeepMind has created a neural network that learns how to play video games in a fashion similar to that of humans.

As opposed to other AIs, such as IBM's Deep Blue or Watson, which were developed for a pre-defined purpose and only function within its scope, DeepMind claims that its system is not pre-programmed: it learns from experience, using only raw pixels as data input.

Primary use-case: TBD

DeepQA (Watson)

Originator: DeepQA Research Team

Status: [state]

Key links: https://researcher.watson.ibm.com/researcher/view_group.php?id=2099

https://en.wikipedia.org/wiki/Watson_(computer)

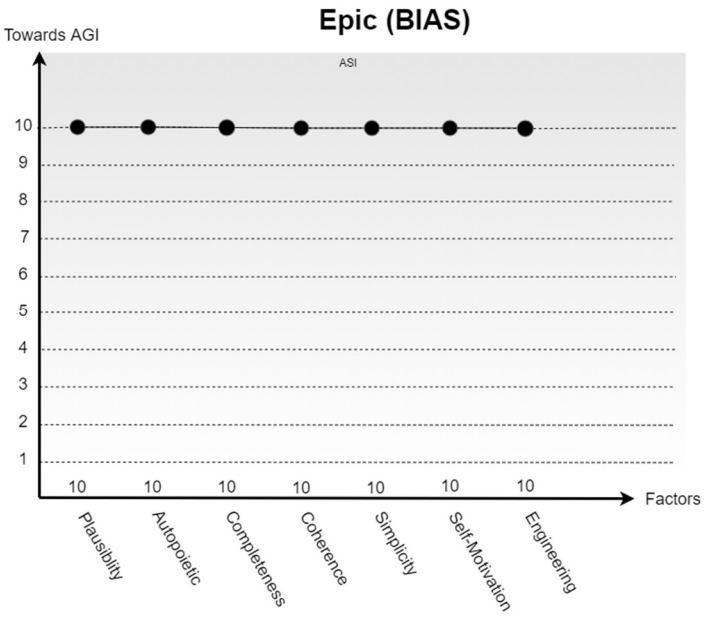
Description: Watson was created by IBM's DeepQA Research Team.

Primary use-case: TBD

EPIC

Originator: Anthony Hornof, David Kieras

Status: Theory



Key links: https://web.eecs.umich.edu/~kieras/epic.html

http://www.umich.edu/~bcalab/epic.html

https://ix.cs.uoregon.edu/~hornof/publications.html

Description: The specific goal is to develop and validate a cognitive modeling architecture called EPIC (Executive-Process/Interactive Control) for human information processing that accurately accounts for the detailed timing of human perceptual, cognitive, and motor activity. EPIC provides a framework for constructing models of human-system interaction that are accurate and detailed enough to be useful for practical design purposes. EPIC represents a state-of-the-art synthesis of results on human perceptual/motor performance, cognitive modeling techniques, and task analysis methodology, implemented in the form of computer simulation software.

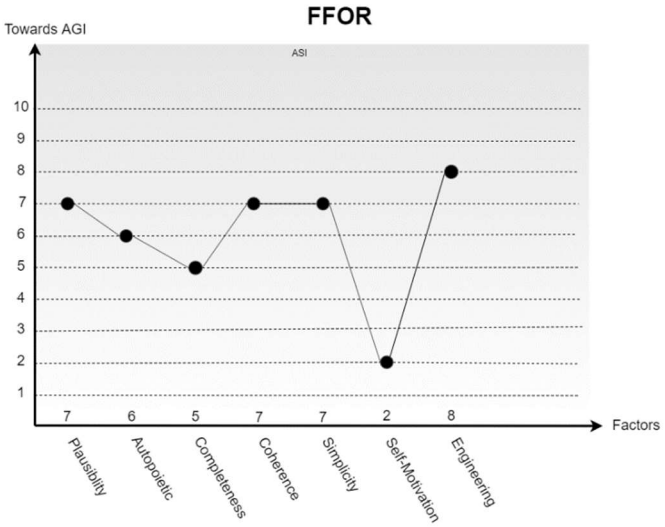
Human performance in a task is simulated by programming the cognitive processor with production rules organized as methods for accomplishing task goals. The EPIC model is then run in interaction with a simulation of the external system and performs the same task as the human operator would. The model generates events (e.g., eye movements, keystrokes, vocal utterances) whose timing is accurately predictive of human performance. The figure shows the overall architecture of an EPIC model interacting with a simulated system. From the links above.

Primary use-case: TBD

FFOR

Originator: Susan L. Epstein,
Department of Computer Science,
Hunter College of The City University of New York,
695 Park Avenue, New York, NY 10065
susan.epstein@hunter.cuny.edu

Status: Code Complete



Key Links: https://en.wikipedia.org/wiki/FORR

http://www.compsci.hunter.cuny.edu/~epstein/html/forr.html

Description: FORR is a cognitively plausible architecture for learning and problem-solving. Its premise is that satisficing based on reactivity, heuristics, planning, and limited search is both human-like and effective. Given a set of domain-general, boundedly rational, user-defined procedures (*Advisors*) and algorithms that learn to instantiate domain-specific concepts, a FORR-based program learns to specialize its behavior to develop problem-class-specific expertise. For example, the game player Hoyle has learned how to play two-person, perfect-information, finite-board games (its domain) as well as the best human competitors [1]. Hoyle knows 19 different games (its problem classes), and its Advisors include one that selects a single step to a win and another that moves to a previously successful state. As a second example, the constraint solver ACE has learned how to solve constraint satisfaction problems (its domain), where a problem class is a set of related constraint problems (e.g., Sudoku puzzles of a certain size) [2]. ACE's Advisors include one that makes a single assignment to a complete solution and another that assigns a value to a variable of maximum dynamic domain size.

Given a set of permissible actions, FORR makes each decision with a pass through a three-tier architecture. Examples here are drawn from SemaFORR, a robot controller for indoor navigation [3]. FORR's first-tier contains pre-ordered reactive Advisors. Some are simple condition-action rules (e.g., "do not move directly into a wall"); others are reactive planners, each of which respond to a situation with a sequence of decisions intended to alleviate it (e.g., "after prolonged presence in a confined space, leave it"). FORR's

second-tier contains deliberative Advisors, planners that formulate a one-time plan to solve a specific problem in a problem class (e.g., "travel efficiently to <*x,y*> in this two-dimensional space"). Because plans are often partial and/or hierarchical, there is also an Advisor in tier 1 that stores the current plan, operationalizes it, and integrates its execution with the reactive components. If FORR has a plan for a task, it typically follows it with tier 1. If, however, the next plan step is impossible (e.g., a door is temporarily obstructed) or several actions would fulfill that step, FORR's third tier makes a choice. Tier-3 Advisors generate heuristic preferences with numerical strengths that support or oppose individual actions (e.g., long steps forward or a small left turn). The output of a FORR decision cycle is either a direct choice made in tier 1 or the winner of a weighted vote that reflects the confidence and reliability of the tier-3 Advisors as a group.

[1] Epstein, S. L. 2001. Learning to Play Expertly: A Tutorial on Hoyle. *Machines That Learn to Play Games*. Fürnkranz, J. and M. Kubat. Huntington, NY, Nova Science**:** 153-178.

[2] Petrovic, S. and S. L. Epstein 2008. Random Subsets Support Learning a Mixture of Heuristics. *International Journal on Artificial Intelligence Tools* 17(3): 501-520.

[3] Epstein, S. L. and R. Korpan 2019. Planning and Communication with a Learned Spatial Model. In Proceedings of 14th International Conference on Spatial Information Theory (COSIT 2019.

[4] Ratterman, M. J. and S. L. Epstein 1995. Skilled like a Person: A Comparison of Human and Computer Game Playing. In Proceedings of Seventeenth Annual Conference of the Cognitive Science Society, 709-714. Pittsburgh, Lawrence Erlbaum Associates.

Primary Use-case: FORR is fully operational and has been coded in Lisp, Java, and C++. It has had multiple primary use cases, including Hoyle, ACE, and SemaFORR. Each of these was extensive work and is a primary use case.

Flowers

Originator: Pierre-Yves Oudeyer and David Filliat

Status: [state]

Key links: https://flowers.inria.fr/

Description: The Flowers project team at Inria, University of Bordeaux, and Ensta ParisTech, studies models of open-ended development and learning. These models are used as tools to help us understand better how children learn, as well as to build machines that learn like children, i.e., developmental artificial intelligence, with applications in educational technologies, automated discovery, robotics, and human-computer interaction.

A major scientific challenge in artificial intelligence and cognitive sciences is to understand how humans and machines can efficiently acquire world models, as well as open and cumulative repertoires of skills over an extended time span.

Primary use-case: TBD

Glair

Originator: Stuart C. Shapiro

Status: [state]

Key links: https://cse.buffalo.edu/~shapiro/Papers/glairPublished.pdf

Description: GLAIR (Grounded Layered Architecture with Integrated Reasoning) is a multilayered cognitive architecture for embodied agents operating in real, virtual, or simulated environments containing other agents.

Primary use-case: TBD

GoodAI

Originator: Marek Rosa

Status: [state]

Key links: https://www.goodai.com/about/

Description: GoodAI was founded in 2014 with a $10M personal investment from Marek Rosa. The long-term goal is to build general artificial intelligence that will automate cognitive processes in science, technology, business, and other fields.

Primary use-case: TBD

Hyperon

Originator: Ben Goertzel

Status: [state]

Key links: https://wiki.opencog.org/w/Hyperon

Description: Hyperon is a substantially revised, novel version of OpenCog -- which is currently (Sep 2020) in an early stage of designing and prototyping.

Among other theoretical foundations, the Hyperon design tries to take the General Theory of General Intelligence into account, at least to some extent. Although the connection with the possible implementation may turn out to be loose, this theory (which is in a moderately advanced but certainly not completed stage) provides some background ideas, principles, and problems constituting the grand picture behind the Hyperon design.

Hyperon is also experimenting with multi-agent control in the Minecraft environment as a simpler use-case for fleshing out and demonstrating Hyperon concepts and capabilities.

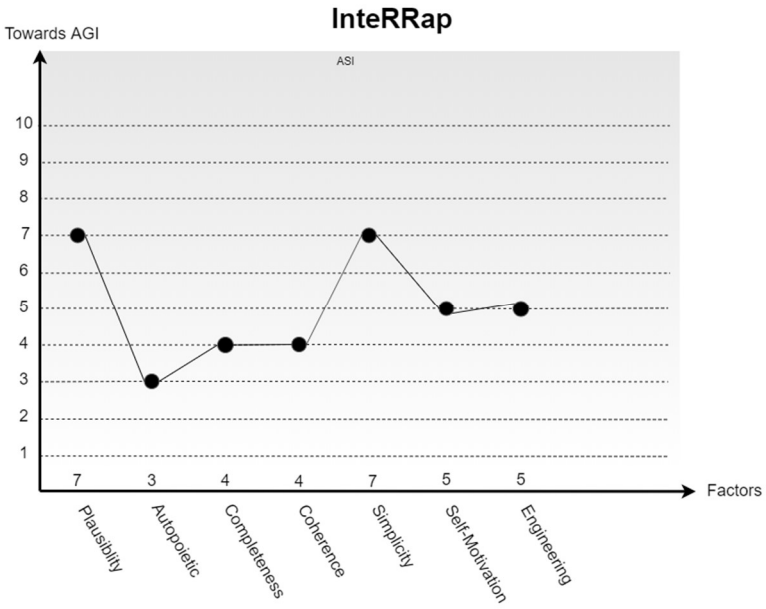
Primary use-case: TBD

InteRRaP

Originator: Jörg Müller

joerg.mueller@tu-clausthal.de

Status: Toy



Key links: https://www.springer.com/de/book/9783540620037
https://www.in.tu-clausthal.de/?id=206
https://meclab.in.tu-clausthal.de/

Description: InteRRaP is a vertically layered cognitive agent architecture. The lowest (behavior-based) layer serves two purposes: first, based on a rule-based runtime model, it enables reactive behavior; second, it enables designers to compile plans into procedural patterns of behavior to efficiently execute routine tasks. The middle (local planning) layer provides local planning from second principles(based on a Belief, Desire, International style plan library) to achieve the agent's goals; the uppermost (cooperative planning) layer aims at enabling planned multi-agent interactions, negotiation, and communication-based conflict resolution and coordination.

All layers have access to a knowledge base, which is vertically partitioned, such that the higher layers have access to additional information (e.g., planning and cooperation). These segments together represent the agent and its environment at different levels of abstraction.

Overall control of an InteRRaP agent is defined through interaction between the layers. It is based on two principles: bottom-up activation and top-down execution. First, agent activity is triggered from "the bottom" (i.e., the behavior-based layer). If a layer is not able to deal with the situation, it passes control to the next-higher layer. For instance, if a situation requires coordination with another agent, control will ultimately reach the cooperative planning level, which may negotiate a joint plan with that agent. Then, the local projection of the joint plan will be passed down to the local planning layer and will be

executed. In this way, control in InteRRaP will flow from the lowest layer to higher layers of the architecture and then back down again.

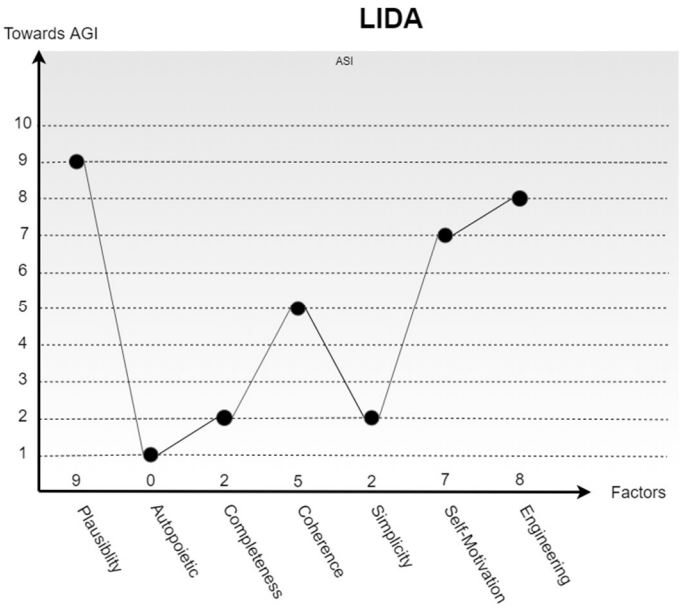
Primary use-case: physical multiagent systems, normative control architecture for IT ecosystems, and modeling and simulation of sociotechnical systems

LIDA (Learning Intelligent Distribution Agent)

Originator: Stan Franklin

franklin@memphis.edu

Status: Complete



Key link: https://ccrg.cs.memphis.edu/papers

Description: LIDA is a systems-level cognitive model. It is conceptual and partly computational. It attempts to model minds, be they human, animal, or artificial, which we take to be control structures for autonomous agents. We think of minds as being implemented as virtual machines running on top of underlying devices such as brains or computers. Every animal must frequently sample its environment, external or internal, and act appropriately in response. The LIDA model's cognitive cycle, taken from the action-perception cycle of the psychologists and neuroscientists, enables just such frequent (~10 Hz in humans) sampling and responding. One can think of the cognitive cycle as a cognitive atom of which higher-level cognitive processes, deliberation, reasoning, problem-solving, planning, imagining, etc., are comprised. Each cognitive cycle can be divided into three phases, a perception and understanding phase, an attention phase, and an action and learning phase. Using incoming sensory data, memories, etc., the first phase updates its understanding of the current situation. The attention phase then filters the content of this understanding for saliency and broadcasts this conscious content globally in accordance with Global Workspace Theory (GWT). Although cognitive cycles may overlap, partially operating in parallel, conscious broadcasts occur in sequence. The third phase selects and executes an appropriate response and also learns into a bevy of memory systems.

Primary use-case:

The first LIDA agent, IDA, found new jobs for US Navy sailors at the end of their current tour of duty, communicating with them via email in natural language and using the Navy's databases. The subsequent

half-dozen or so LIDA agents all replicate subjects in psychological experiments. These LIDA agents serve to verify aspects of the theory.

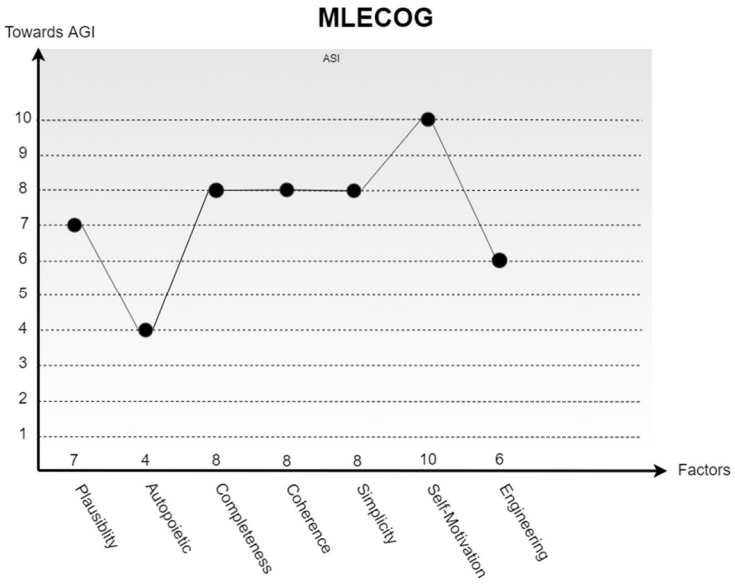
Building a Better Humanity

MLECOG

Originator: Professor Janusz Starzyk,

University of Information Technology and Management, Rzeszow, Poland.

Status: Partial Implementation



Key links: http://ncn.wsiz.rzeszow.pl/

http://perception.wsiz.rzeszow.pl/

Description: Also referred to as Motivated Embodied Mind, the (MEM)MLECOG is a cognitive architecture designed to be a brain of an embodied agent (real or virtual) and aims at achieving general artificial intelligence. The agent develops its skills and knowledge by acting on its environment and observing the results of its actions. Its motivations, goal creation, planning, and thinking are based on the principles of motivating learning that use a reduction of pain signals as a foundation for learning. Besides primary goals that are set by the designer, the agent creates its internal goals and receives an internal reward for attaining these goals.

MLECOG aims at achieving consciousness in the machine with pain signals and emotional states that create feelings and drive its behavior. It uses visual saccades, attention, and attention switching to focus, observe and act. Mental saccades drive its working memory to think, plan and imagine. Its knowledge will grow gradually through observations, learning of new skills, episodic development, cooperation, and teacher input.

Key use-cases: Not sure how to answer this question or even if I understand it correctly. As described in point 4. the architecture is partially implemented with the code describing motivated learning agents in a virtual world. Currently, we concentrate on the integration of perception, semantic and episodic memory with attention switching and the motivated learning value system.

NARS (Non-Axiomatic Reasoning System)

Originator: Pei Wang

Status: Theory

[valence diagram]

Key links: https://cis.temple.edu/~pwang/NARS-Intro.html

Description: NARS (Non-Axiomatic Reasoning System) is a project aimed at the building of a general-purpose intelligent system, i.e., a "thinking machine" (also known as "AGI"), that follows the same principles as the human mind and can solve problems in various domains.

The design of NARS is based on the belief that the essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real-time, be open to unexpected tasks, and learn from experience.

To realize this form of intelligence in a computer, NARS takes a unified approach; that is, the system depends on a single core technique to carry out various cognitive functions and to solve various problems.

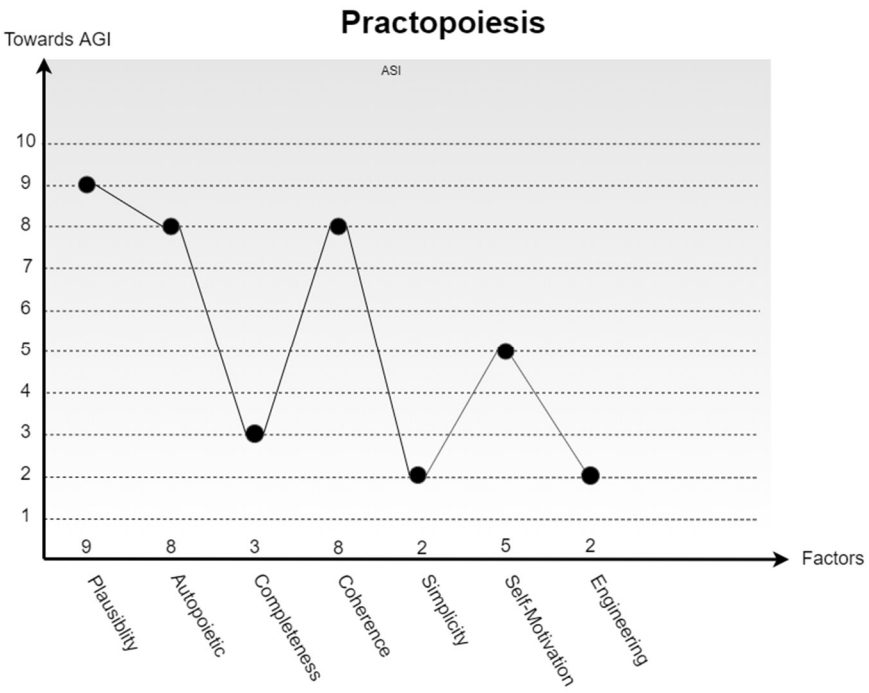
Primary use-case:

AGI - Already integrated into a Cisco framework that has been used in traffic monitoring, inventory management, and network event analysis.
See "A Reasoning Based Model for Anomaly Detection in the Smart City Domain," By Patrick Hammer, Tony Lofthouse, Enzo Fenoglio, Hugo Latapie, and Pei Wang
Proceedings of IntelliSys 2020, Pages 144-159, Online, September 2020

Practopoiesis (aka, hierarchical adaptations)

Originator: Danko Nikolic

Status: Theory



Key links: www.danko-nikolic.com/practopoiesis

http://www.robotsgomental.com/

http://www.ai-kindergarten.com/

Description: The system does not only rely on the internal computation of a neural network, but it uses its environment in addition--the situation in which it finds itself. The sensory inputs coming from the environment help in the form of a fast adaptation to the surrounding situation. The intelligence works such that the knowledge acquired in the past is combined with the sensory inputs in order to create a new neural network for every new situation in which the AI finds itself. Traditional neural networks "think" by running activations through the existing neural networks. AI that is based on practopoiesis "thinks" by rewiring a neural network. The rewiring process corresponds to its mental activity.

Primary use-case: A home robot

Recommendation Architecture

Originator: L. Andrew Coward

landrewcoward@gmail.com

Status: Toy

[valence diagram]

Key links: https://dl.acm.org/doi/abs/10.1016/S1389-0417%2801%2900024-9
www.springer.com/us/book/9789400771062

Description: The primary focus of work on the Recommendation Architecture is understanding higher cognition in terms of neuron mechanisms. The connection with AGI is a set of theoretical arguments that there are strong constraints on the architecture of any system which needs to learn a complex combination of different types of behaviors. These constraints derive from the need to avoid excessive information processing resources and the need to learn new behaviors without severe interference with earlier learning. As the ratio of behaviors to resources increases, a learning system is more and more tightly constrained into the recommendation architecture. Hence a sufficiently general AGI system, like the brain, will almost certainly be constrained into the recommendation architecture.

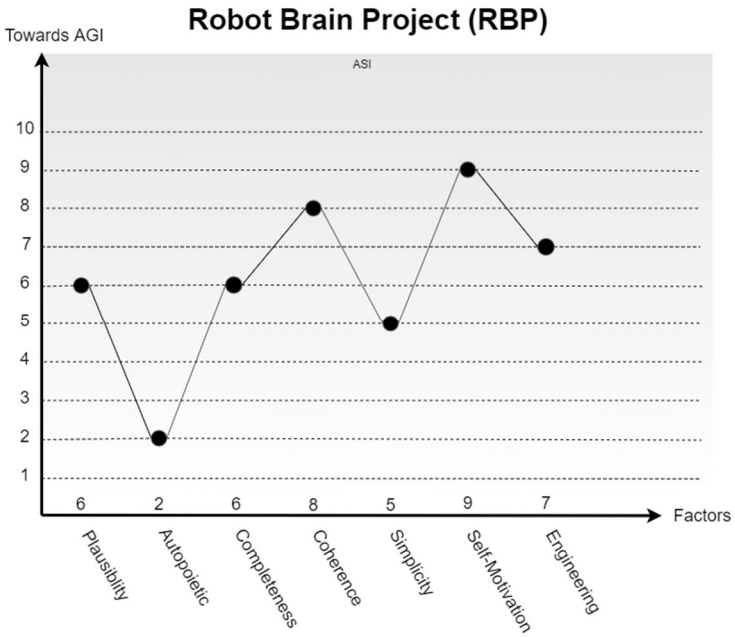
The brain has subsystems and sub-subsystems etc., that closely resemble the subsystems of this recommendation architecture as defined by the theoretical arguments. Major brain subsystems include the cortex, hippocampus, basal ganglia, thalamus, amygdala, and cerebellum. The architecture at this level of major subsystems is illustrated. The cortex defines and detects conditions within the information available to the brain. The hippocampus manages changes to cortical conditions. The basal ganglia get current condition detections from the cortex and interpret each detection as a set of recommendations in favor of a wide range of behaviors, each with an individual recommendation weight. The basal ganglia determine and select the most strongly recommended behavior(s). Some cortical conditions detected following behavior can recommend rewards (positive or negative). This reward feedback cannot guide the definition of conditions; it can only modify the recommendation weights in favor of recent behaviors. Most behaviors are implemented by the release of condition detections into, within, or out of the cortex, and the thalamus implements these behaviors after they are selected by the basal ganglia. The amygdala biases the determination of behavior in favor of one of a number of general types (aggressive, fearful, etc.). For sequences of behaviors that have been previously learned and often used, the cerebellum takes over the implementation. This saves the time required in the basal ganglia for the selection of each individual behavior, so the execution of the sequence is faster. However, for any changes to behaviors or to their order, control must go back to the basal ganglia.

Primary use-case: Human Brain Modeling

Robot Brain Project (RBP), formerly BECCA (Brain Emulating Cognitive Control Architecture)

Originators: Brandon Rohrer

Status: Code Complete



Key links: https://e2eml.school/robot_brain_project.html

Description: The Robot Brain Project is an effort to make a plug-and-play robot brain, a general reinforcement learning solution that is well suited to physical robots pursuing arbitrary goals in physical environments. Superficially, it bears some similarity to existing deep reinforcement learning techniques. It has a hierarchical feature-learning component (analogous to a neural network) and a model-based policy learning component (analogous to existing RL algorithms). RBP differs in a few important ways:

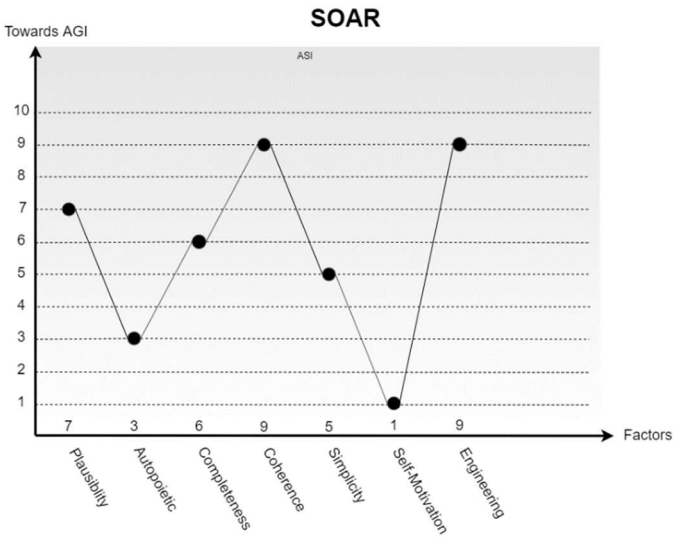
- It doesn't assume that inputs are of the same type (e.g., all pixels or all numeric)
- It doesn't assume that inputs are arrayed.
- All inputs are discretized.
- Feature learning is unsupervised (it doesn't rely on backpropagation and so doesn't need to be differentiable).
- It has "life-long learning," that is, it doesn't have distinct training and testing phases.
- It grows as many layers as the data supports. It can build features of arbitrary complexity.
- It builds a large parallel set of feature-by-feature transition models rather than a full state-action-state transition model.
- It doesn't assume a stationary world or even a stationary physical embodiment.

Key use-cases: RBP will be successful when it can be plugged into any robot hardware, having any set of rewards, and learning to behave in a way to get those rewards. It won't need to know anything about the environment, the task, or its own physical embodiment beforehand.

Soar

Originated by: John Laird

Status: Soar is used for a variety of robot platforms, synthetic characters in simulation/games, decision support systems, autonomous vehicles, training systems.



Key links: https://soar.eecs.umich.edu/

Description: Soar is a general cognitive architecture for developing systems that exhibit intelligent behavior.
We intend ultimately to enable the Soar architecture to:
• work on the full range of tasks expected of an intelligent agent, from highly routine to extremely difficult, open-ended problems
• represent and use appropriate forms and varying levels of knowledge, such as procedural, semantic, episodic, and iconic
• employ the full range of problem-solving methods
• interact with the outside world, and
• learn about all aspects of the tasks and their performance on them.
In other words, our intention is for Soar to support all the capabilities required of a general intelligent agent.

Key use-cases: Soar is used for a variety of robot platforms, synthetic characters in simulation/games, decision support systems, autonomous vehicles, training systems.

Sigma

Originators: Prof. Paul Rosenbloom

rosenbloom@usc.edu

Dr. Volkan Ustun

ustun@ict.usc.edu

Status: Sigma is currently implemented in Lisp, with both publicly available and development versions, but with experimental ports to other languages also being investigated for improved speed and portability.

[valence diagram]

Key links: https://cogarch.ict.usc.edu/

https://sites.usc.edu/rosenbloom/

https://ict.usc.edu/profile/volkan-ustun/

Description: Sigma is a non-modular, hybrid (continuous + discrete), mixed (symbolic + probabilistic/activation) architecture that is being developed in service of four key desiderata: (1) grand unification, spanning not only traditional cognitive capabilities but also key non-cognitive aspects; (2) generic cognition, spanning both natural and artificial cognition; (3) functional elegance, yielding the full set of capabilities necessary for human(-level) intelligence from a simple and theoretically elegant base; and (4) sufficient efficiency, executing quickly enough for real-time applications and large-scale experiments. The architecture is defined across two layers, the lower one originally grounded in the elegant generality of graphical models but since minimally extended to handle rules, sum-product networks, and neural networks; and the upper one defining cognitive memories, representations, and processes through defining, solving and modifying these graphs. Processing is driven by a cognitive cycle that includes a parallel elaboration phase for graph solution followed by an adaptation phase for decision making and learning.

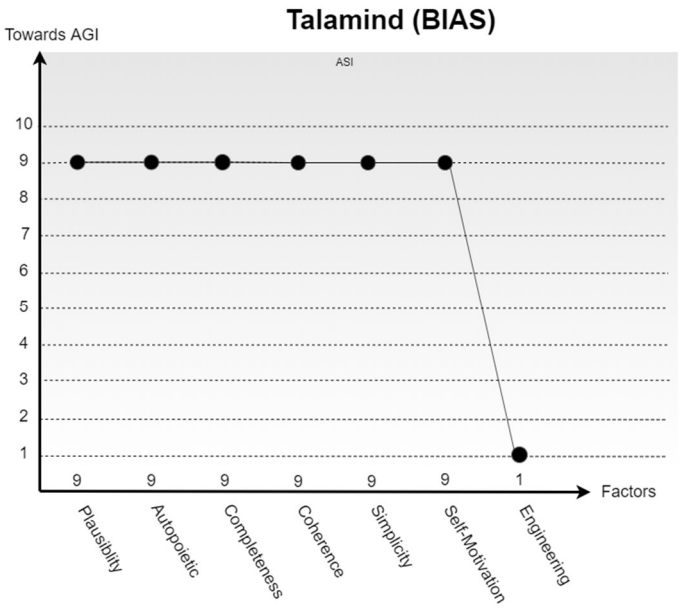
Based on these two layers, Sigma has been able to demonstrate a variety of forms of memory (e.g., procedural, declarative, constraint, perceptual, distributed vector, and neural network), learning (e.g., concept, episodic, reinforcement, action modeling, models of other agents, perceptual, and backpropagation), problem-solving (e.g., preference-based, motivated, and decision-theoretic decision making; impasse-driven reflection; and Theory of Mind reasoning), mental imagery, perception (e.g., object and speech recognition, and spatial localization), natural language (e.g., word sense disambiguation, part of speech tagging, sentence identification, and dialogue), emotion, and motivation. Demonstrations also exist of combinations of these various capabilities in the context of both toy domains and several more realistic virtual/simulated human tasks. Work has also begun on additional capabilities, such as personality and conversational question answering.

Primary use-case: The primary use-case is the minds of virtual humans and other forms of human-like intelligent agents for complex simulation environments. In particular, the focus has been on working towards applications in domains that require human-like autonomous social cognitive (HASC) systems.

TalaMind

Originator: Phil Jackson

Status: Theory



Key links: https://www.talamind.com

https://www.researchgate.net/profile/Philip_Jackson_Jr_Phd

Description: 'TalaMind' is a research project, theoretical approach, and systems architecture for eventually achieving human-level artificial intelligence. It was proposed in a doctoral thesis (Jackson, 2014) and has been further discussed in subsequent papers and books, listed below. The TalaMind architecture has a linguistic, archetype, and associative 'levels' – each level can have additional structure and layers within it. At the linguistic level, the architecture includes a 'natural language of thought' called Tala, a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual framework to produce intelligent behaviors and new concepts. The archetype level is where cognitive categories are represented using methods such as conceptual spaces, image schemas, radial categories, etc. The associative level would typically interface with a real-world environment and support connectionism (deep neural networks), Bayesian processing, etc.

To date, work on the TalaMind approach has been agnostic about research choices at the archetype and associative levels. The architecture is open to design choices at the three conceptual levels, for instance, permitting predicate calculus, conceptual graphs, and other symbolisms in addition to the Tala language at the linguistic level, and permitting integration across the three levels, e.g., the potential use of deep neural networks at the linguistic and archetype levels. Because it envisions combining symbolic processing and connectionism in a hybrid architecture, TalaMind may be considered a 'neuro-symbolic' research approach toward eventually achieving human-level AI.

Building a Better Humanity

Primary use-case: Human Level AGI

Tencent AI

Originator: [name]

Status: [state]

Key links: https://ai.qq.com/hr/ailab.shtml

https://ai.tencent.com/

Description: AI Lab intent on 'Making AI commonplace.'

Primary use-case: TBD

Vicarious

Originator: [name]

Status: [state]

Key links: https://www.vicarious.com/science/

Description: An AI and Robotics company pride itself on backing from Bezos, Musk, Zuckerburg, etc.

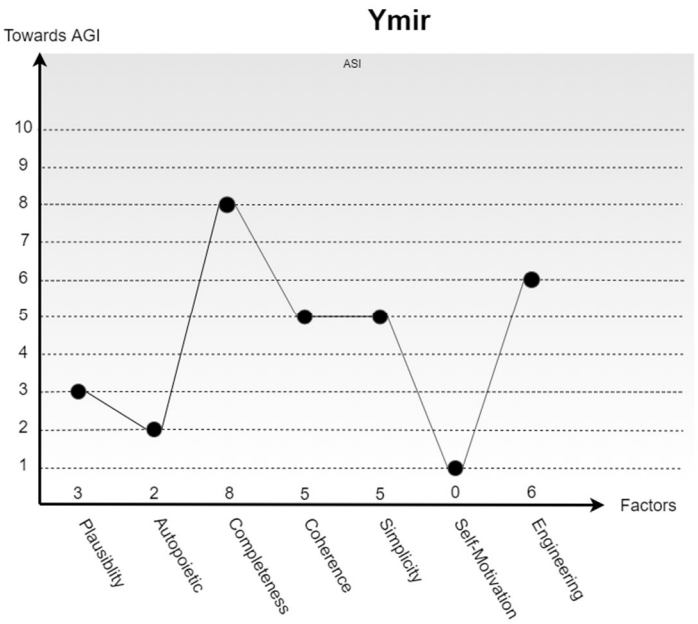
Aims to make robotics and automation commonplace in 15 years, with AGI a long-term goal.

Primary use-case: TBD

Ymir

Originator: Kristinn R. Thórisson

Status: Toy



Key links: http://alumni.media.mit.edu/~kris/ymir.html

Description: Ymir was the Ph.D. thesis of Kristinn R. Thórisson at the MIT Media Lab. Ymir is a generative model of human psychosocial dialogue skills. Instead of dealing with a single issue, or a few aspects of human face-to-face multimodal dialogue, Ymir addresses all key issues needed to create artificial characters that can engage in real-time situated dialogue with humans. Ymir does resource-bounded problem solving where the problem is dialogue; the resources are time, information, and computational power. Ymir has been used to create softbots (and robots) whose purpose in life is to receive commands from humans, ask questions when appropriate, and otherwise turn commands and instructions into executable action in their domain of expertise. At the time of its construction in the early 90s, Ymir was used to test theories about human discourse, as it provides the possibility to turn certain dialogue actions and skills on and off at will—something that was impossible to do before, even with a skilled actor. Many of the features of Gandalf — the first agent built in the architecture — are still unparalleled in state-of-the-art dialog systems, including its fine-grained gaze control, finely-coordinated manual gesture and speech generation, and real-time automatic prosody analysis to predict end-of-sentence of the interlocutor at a high temporal resolution. The design of Ymir, like the subsumption architecture and most good-old-fashioned AI architectures of the last century, follows a constructionist paradigm, where numerous hand-crafted rules and coordination mechanisms work in tandem to create a coordinated whole. Ymir consists of these building blocks: 1. A set of semi-independent processing layers, G.2. A set of blackboards, F.4. A set of perceptual modules, r. 5. A set of decision modules, P. 6. A set of behaviors, b, and behavior morphologies, bm (specific motor programs). 7. A set of knowledge bases, k. The main distinguishing features of Ymir are 1. A distributed, modular approach to perception, decision, and action. 2. A layered combination of reactive and reflective behaviors. 3. Dialogue-related interpretation is separated from topic interpretation. 4. Dialogue management is viewed as having complete process control (when something happens as opposed to what happens) of overt and covert

actionn.5. Motor actions are split into two phases; a decision (or intentional) phase and a composition/execution phase.6. Intentions to act vary in their specificity: the more specific an intention is (e.g., blinking), the fewer morphologies (ways to do it) exist; the less specific it is (e.g., looking confused), the more options there are in the way it will eventually be realized.7. The final morphology of an intention is chosen at runtime.8. The architecture does not and was not intended to learn.

Primary use-case: Situated real-time dialog control for embodied humanoid agents.

Appendix E: Bibliography

Ahmed, H.; Glasgow, J.; "Swarm Intelligence: Concepts, Models and Applications"; School of Computing, Queen's University; Feb 2013

Altevogt, B., Pankevich, D., Shelton-Davenport, M., Kahn, J., (2009), "Chimpanzees in biomedical and behavioral research: Assessing the Necessity," Washington, DC: *The National Academies Press. Institute of Medicine (US) and National Research Council (US) Committee on the Use of Chimpanzees in Biomedical and Behavioral Research*; Washington (DC); IOM (Institute of Medicine); National Academies Press (US), 2011, ISBN-13: 978-0-309-22039-2ISBN-10: 0-309-22039-4 https://www.ncbi.nlm.nih.gov/books/NBK91445/

Agrawal, P.; "M25 – Wisdom"; Speakingtree. in – 2017 - http://www.speakingtree.in/blog/m25wisdom

Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang. (2019) "Speech synthesis from neural decoding of spoken sentences." Nature 568 (7753): 493–498.

Azure: Common web application architectures; Accessed Feb 2021; Microsoft 2020 https://docs.microsoft.com/en-us/dotnet/architecture/modern-web-apps-azure/common-web-application-architectures

Babcock, J.; Kramar, J.; Yampolskiy, R.; "The AGI Containment Problem" arXiv: 1604.00545; Cornell University; DOI; 10.1007/978-3-319-41649-6;

Baars, B.; "Current concepts of consciousness with some implications for anesthesia;" Refresher Course Online – Canadian Anesthesiologists Society 2003; The Neurosciences Institute, San Diego CA

Baars, B.; "Subjective Experience is probably not limited to humans: The evidence from neurobiology and behavior;" The Neurosciences Institute, San Diego; 2004 Elsevier

Baars, B.; "The Global Workspace Theory of Consciousness – Prediction and Results;" The Blackwell Companion to Consciousness, Second Edition, 17 Mar 2017; https://doi.org/10.1002/9781119132363.ch16

Baars, B.; Katherine, M; Global Workspace; 28 NOV 2016; UCLA http://cogweb.ucla.edu/CogSci/GWorkspace.html

Baars, B.; McGovern, K.; "Global Workspace – A Theory of Consciousness;" November 5, 1997; UCLA Berkeley, California; Last used: http://cogweb.ucla.edu/CogSci/GWorkspace.html

Barrat, J.; "Our Final Invention – Artificial Intelligence and the end of the Human Era;" Thomas Dune Books; St. Martin's Griffin, NY; 2013; ISBN: 978-0-312-62237-4

Barrett, L.; "How Emotions Are Made – The Secret Life of the Brain"; Houghton Mifflin Harcourt (March 7, 2017); ISBN-10: 9780544133310

Barrett, L. F.; Tugade, M. M.; Engle, R. W. (2004). "Individual differences in working memory capacity and dual-process theories of the mind." Psychological Bulletin. 130 (4): 553–573. doi:10.1037/0033-2909.130.4.553. PMC 1351135. PMID 15250813.

Beavers, A. “Alan Turing: Mathematical Mechanist.” 2013, In Cooper, S. Barry; van Leeuwen, Jan (eds.). Alan Turing: His Work and Impact. Waltham: Elsevier. Pp. 481–485. ISBN 978-0-12-386980-7.

Bostrom, N.; Ryan, N.; et al.; “Superintelligence – Paths, Dangers, Strategies;” Oxford University Press; 2014, ISBN-13: 978-019968112; ISBN-10: 0199678111;

Camp, Jim; “Decisions Are Emotional, Not Logical: The Neuroscience behind Decision Making;” 2016 http://bigthink.com/experts-corner/decisions-are-emotional-not-logical-the-neuroscience-behind-decision-making

Chakraborty, A.; Kar, A.; “Swarm Intelligence: A Review of Algorithms”; Springer International Publishing AG 2017 DOI 10.1007/978-3-319-50920-4_19

Chalmers, D.; Facing Up to the Problem of Consciousness; University of Arizona 1995

Chollet, F.; “On the Measure of Intelligence;” arXiv:1911.01547; https://arxiv.org/abs/1911.01547

Clarke, A.; “Hazards of Prophecy: The Failure of Imagination,” in Profiles of the Future: An Enquiry into the Limits of the Possible, Pan Books, pp. 14, 21, 36; 1973

Coyle, D.; “The Culture Code – The Secrets of Highly Successful Groups”; Bantam 2018; ISBN-13: 978-0304176989

CRASSH (2016) A symposium on the technological displacement of white-collar employment: political and social implications.”; Wolfson Hall, Churchill College, Cambridge

Crevier, D.; “AI: The Tumultuous Search for Artificial Intelligence;” 1993; New York, NY: BasicBooks, ISBN 0-465-02997-3

Crosby, M.; Beyret, B.; Shanahan, M.; Hernandez-Orallo, J.; Cheke, L.; Halina, M.; “The Animal-AI Testbed and Competition;” Proceedings of Machine Learning Research 2020; NeurIPS Competitions and Demonstrations; https://www.mdcrosby.com/Animal_AI.pdf

Damasio, A.; “This Time with Feeling: David Brooks and Antonio Damasio;” Aspen Institute 2009; https://www.youtube.com/watch?v=IifXMd26gWE

Damasio, A.; “The feeling of what happens: body and emotion in the making of consciousness;” Harcourt Brace; 1999

Dambrot, S. Mason. (2016) “Exocortical Cognition: Heads in the Cloud - A transdisciplinary framework for augmenting human high-level cognitive processes.” 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, pp. 004007–004014.

Dienes, Z; Seth, A.; The conscious and unconscious; University of Sussex; 2012

Dienes, Z; Seth, A.; Measuring any conscious content versus measuring the relevant conscious content: Comment on Sandberg et al.; Elsevier/ScienceDirect; University of Sussex 2010

Einstein, A/.; “Cosmic Religion: With Other Opinions and Aphorisms.” Dover Publications, p. 97.; 1931

Engel, D.; Woolley, A.; Chabris, C.; Takahashi, M.; Aggarwal, I.; Nemoto, K.; Kaiser, C.; Kim, Y.; Malone, T.; "Collective Intelligence in Computer-Mediated Collaboration Emerges in Different Contexts and Cultures;" Bridging Communications; CHI 2015; Seoul Korea

Everitt, T.; Lea, G.; Hutter, M.; "AGI Safety Literature Review;" In International Joint Conference on Artificial Intelligence (IJCAI). ArXiv: 1805.01109.

Gans, J. "AI and the paperclip problem;" 10 Jun 2018; VOX EU CEPR; Accessed 30 Sept 2020; https://voxeu.org/article/ai-and-paperclip-problem

Goertzel, B.; Wigmore, J.; "The Puzzle: Why is it so hard to Measure Partial Progress Toward Human-Level AGI?;" Blog: "The Multiverse According to Ben;" Accessed 11 Jan 2022; http://multiverseaccordingtoben.blogspot.com/2011/06/why-is-evaluating-partial-progress.html

Gill, K.; "Artificial Super Intelligence: Beyond Rhetoric"; Springer-Velage London 2016; Feb 2016; AI & Soc. (2016) 31:137-143; DOI 10.1007/s00146-016-0651-x

Google; "General AI challenge;" Accessed 12 Jan 2022; https://www.general-ai-challenge.org/

Graves, A.; Abdel-rahman Mohamed, and Geoffrey Hinton. (2013) "Speech recognition with deep recurrent neural networks." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC: 6645–6649.

Gregory; "Qualia: What it is like to have an experience; NYU; 2004 https://www.nyu.edu/gsas/dept/philo/faculty/block/papers/qualiagregory.pdf

Hawkins, J.; "On Intelligence"; Times Books; Adapted edition (October 3, 2004); ISBN-10: 0805074562

Hawryszkiewycz, I.; "Introduction to system analysis and design;" Prentice Hall PTR, 1994

Heath, C.; Larrick, R.; Klayman, J.; "Cognitive Repairs: How Organizational Practices Can Compensate For Individual Short Comings"; Research in Organizational Behavior Volume 20, pages 1-37; ISBN: 0-7623-0366-2

Hernández-Orallo, J.; Dowe, D.; "Measuring universal intelligence: Towards an anytime intelligence test;" Artificial Intelligence; 174 (2010) 1508–1539],

Hu, Y.; "Swarm Intelligence"; (presentation)

Iphigenie; "What are the differences between sentience, consciousness, and awareness?"; Philosophy – Stack Exchange; https://philosophy.stackexchange.com/questions/4682/what-arethe-differences-between-sentience-consciousness-and-awareness; 2017

Institute for Creative Technologies (ICT); "Cognitive Architecture;" University of Southern California (USC);

James, I.; "Claude Elwood Shannon 30 Apr 1916 – 24 Feb 2001". 2009; Biographical Memoirs of Fellows of the Royal Society. 55: 257–265. DOI:10.1098/rsbm.2009.0015.

Jangra, A.; Awasthi, A.; Bhatia, V.; "A Study on Swarm Artificial Intelligence;" IJARCSSE v3 #8 August 2013; ISSN: 227 128X

Kaufman, Alan S.; Lichtenberger, Elizabeth (2006). Assessing Adolescent and Adult Intelligence (3rd ed.). Hoboken (NJ): Wiley. p. 7. ISBN 978-0-471-73553-3. Lay summary (22 August 2010).

Kelley, D.; "The Independent Core Observer Model Theory of Consciousness and the Mathematical model for Subjective Experience;" By ICIST 2018 – International Conference on Information Science and Technology – China – April 20-22nd. (IEEE conference);Year: 2018, Volume: 1, Pages: 396-400; ISBN: 978-1-5386-6956-3; https://www.computer.org/csdl/proceedings-article/icnisc/2018/695600a396/1dUo3atoEAo

Kelley, D.; "The Independent Core Observer Model Computational Theory of Consciousness and Mathematical model for Subjective Experience"; ITSC 2018; China

Kelley, D.; "The Sapient and Sentient Intelligence Value Argument (SSIVA) Ethical Model Theory for Artificial General Intelligence"; Springer 2019; Book Titled: "Transhumanist Handbook" page: 175; ISBN: 978-3-030-16919-0

Kelley, D.; "Abstract Theory of Consciousness (ATC);" BICA @IS4SI 2021; MDPI Conference Proceedings; Switzerland (pending)

Kelley, D.; "Independent Core Observer Model Research Program Assumption Codex;" BICA 2019, Pre-conference Proceedings: https://www.springer.com/us/book/9783030257187

Kelley, D.; "Architectural Overview of a 'Mediated' Artificial Super Intelligence Systems based on the Independent Core Observer Model Cognitive Architecture;" (pending 2020) BICA;

Kelley, D.; "Critical Nature of Emotions in Artificial General intelligence – Key Nature of AGI Behavior and Behavioral Tuning in the Independent Core Observer Model Architecture Based System;" IEET Institute for Ethics and Emerging Technologies 2016; https://archive.ieet.org/articles/Kelley20160923.html

Kelley, D.; Chapter: "The Intelligence Value Argument and Effects on Regulating Autonomous Artificial Intelligence;" from Book "The Transhumanist Handbook"; Edited by Newton Lee; Springer 2019

Kelley, D.; "Preliminary Results and Analysis Independent Core Observer Model (ICOM) Cognitive Architecture in a Mediated Artificial Super Intelligence (mASI) System;" BICA 2019, Pre-conference Proceedings: https://www.springer.com/us/book/9783030257187

Kelley, D.; "Self-Motivating Computational System Cognitive Architecture: An Introduction" Chapter 24 from "Google It – Total Information Awareness" Edited by Lee, N.; Part VI; Springer, ISBN 978-1-4939-6415-4; http://www.springer.com/us/book/9781493964130; http://www.amazon.com/Google-Information-Awareness-Newton-Lee/dp/1493964135/

Kelley, D.; "The Intelligence Value Argument and Effects on Regulating Autonomous Artificial Intelligence"; Springer 2018

Kelley, D.; "Human-like Emotional Responses in a Simplified Independent Core Observer Model System;" BICA 02017; Procedia Computer Science; https://www.sciencedirect.com/science/article/pii/S1877050918300358

Kelley, D.; "The Human Mind vs. The Independent Core Observer Model (ICOM) Cognitive Architecture;" [Diagram] 19 Mar 2019; ResearchGate; DOI: 10.13140/RG.2.2.29694.64321; https://www.researchgate.net/publication/331889517_The_Human_Mind_Vs_The_Independent_Core_Observer_Model_Cognitive_Architecture

Kelley, D.; [3 chapters] "Artificial General Intelligence and ICOM;" [Book] Google It – Total Information Awareness" By Newton Lee; Springer (ISBN 978-1-4939-6415-4)

Kelley, D.; Atreides, K.; "The AGI Protocol for the Ethical Treatment of Artificial General Intelligence Systems;" Biologically Inspired Cognitive Architectures 2019; Pending Elsevier/Procedia; DOI: 10.13140/RG.2.2.16413.67044

Kelley, D., Twyman, M.A.; "Independent Core Observer Model (ICOM) Theory of Consciousness as Implemented in the ICOM Cognitive Architecture and associated Consciousness Measures," AAAI Sprint Symposia; Stanford CA; Mar.02019; http://ceur-ws.org/Vol-2287/paper33.pdf

Kelley, D.; Twyman, M.; "Approaching the Psychology of Artificial Intelligence;" Elsevier/Procedia Computer Science; Seattle, WA; Proceeding of BICA Society Conference 2020 JVTR;

Kelley, D.; Twyman, M.; "Biasing in an Independent Core Observer Model Artificial General Intelligence Cognitive Architecture" AAAI Spring Symposia 2019; Stanford University

Kelley, D.; Twyman, M.S.; Dambrot, S.M.; "Preliminary Mediated Artificial Superintelligence Study, Experimental Framework, and Definitions for an Independent Core Observer Model Cognitive Architecture based System;" Springer; Biologically Inspired Cognitive Architectures 2019; Proceedings of the Tenth Annual Meeting of the BICA Society; Editors: Samsonovich, Alexei V. (Ed.)

Kelley D., Waser M.; "Feasibility Study and Practical Applications Using Independent Core Observer Model AGI Systems for Behavioral Modification in Recalcitrant Populations.;" In Samsonovich A. (eds) "Biologically Inspired Cognitive Architectures 2018.;" BICA 2018.; 24 Aug 2018; ISBN: 978-3-319-99315-7; pp 165-173; Advances in Intelligent Systems and Computing, vol 848. Springer, Cham. https://doi.org/10.1007/978-3-319-99316-4_22

Kelley, D.; Waser, M.; "Human-like Emotional Responses in a Simplified Independent Core Observer Model System;" BICA 2017 Conference Proceedings, Elsevier, Procedia Computer Science, Science Direct; 123 (2018) 221–227

Knight, H. "Early Artificial Intelligence Projects – A Student Perspective;" Part of NSF's Recovering MIT's AI Film History Project; 2006; Accessed 2 AUG 2020; URL: https://projects.csail.mit.edu/films/aifilms/AIFilms.html#:~:text=The%20first%20coordinated%20AI%20research,26%20and%20the%20Computation%20Center

Kose, U.; Arslan, A.; "On the Idea of a New Artificial Intelligence Based Optimization Algorithm Inspired from the Nature of Vortex";

Kutsenok, A.; "Swarm AI: A General-Purpose Swarm Intelligence Technique"; Department of Computer Science and Engineering; Michigan State University, East Lansing, MI 48825

Kurzweil, R.; The Law of Accelerating Returns; Mar 2001; http://www.kurzweilai.net/the-law-of-accelerating-returns

Kurzweil, R.; Lane, C.; "How to Create a Mind: The Secret of Human Thought Revealed;" ISBN-13: 978-0670025299; Viking; 2012

Leahu, L.; Schwenk, S.; Sengers, P.; "Subjective Objectivity: Negotiating Emotional Meaning;" Cornell University; http://www.cs.cornell.edu/~lleahu/DISBIO.pdf

Legg, S; Hutter, M.; "Universal intelligence: A definition of machine intelligence;" Minds and Machines, 17(4):391-44, 2007

Legg, S; Hutter, M.; "An Approximation of the Universal Intelligence Measure;"

Lim, Hyang-Tag, Jong-Chan Lee, Kang-hee Hong, and Yoon-Ho Kim: (2015) "Avoiding sudden entanglement death using quantum measurement reversal on single-qubit," in CLEO: 2015, OSA Technical Digest (online), Optical Society of America, JW2A.3, pp. 1–2.

Malik, Y.; "Artificial Intelligence and the Hard Problem of Consciousness"; Futuremonger.com; 12 FEB 2018; Accessed 6 Dec 2019; https://futuremonger.com/artificial-intelligence-and-the-hard-problems-of-consciousness-yogesh-malik-d5b63a631627

Malone, T.; "Superminds – The surprising Power of People and Computers Thinking together;" Little Brown and Company, NY, NY; 2018; ISBN-10: 0316349135

Manoogian, J.; Benson, B.; "Cognitive Bias Codex;" Wikipedia; Accessed 2021; https://en.wikipedia.org/wiki/List_of_cognitive_biases

McCarthy, J.; "Professor John McCarthy;" Stanford University 2012; Accessed 30 Sept 2020; http://jmc.stanford.edu/

McCorduck, Pamela (2004), Machines Who Think (2nd ed.), Natick, MA: A. K. Peters, Ltd., ISBN 978-1-56881-205-2, OCLC 52197627

Merriam-Webster—Definition of Consciousness by Merriam-Webster - https://www.merriam-webster.com/dictionary/consciousness

Minsky, Marvin Lee (1986). The Society of Mind. New York: Simon and Schuster. ISBN 978-0-671-60740-1. The first comprehensive description of the Society of Mind theory of intellectual structure and development. See also The Society of Mind (CD-ROM version), Voyager, 1996.

Minsky, M.; Papert, S.; "Perceptron's: An Introduction to Computational Geometry;" 1969; ISBN 0262130432;

Minsky, M.; "The Emotion Machine;" 2006; Simon & Schuster. ISBN 0-7432-7663-9.

MIT; "MIT Center for Collective Intelligence;" Accessed 13 Oct 2020 at https://cci.mit.edu/

Muller, V.; Bostrom, N.; "Future Progress in Artificial Intelligence: A Survey of Expert Opinion"; Synthese Library; Berline: Springer 2014

Neisser, Ulrich (1997). "Rising Scores on Intelligence Tests." American Scientist. 85 (5): 440–447. Bibcode:1997AmSci..85..440N. Archived from the original on 4 November 2016. Retrieved 1 December 2017.

Neto, M.D.C.; Santo, A.; "Emerging Collective Intelligence Business Models;" MCIS 2012; Semantic Scholar; https://www.semanticscholar.org/paper/Emerging-Collective-Intelligence-Business-Models-Neto-Santo/5a6a7780bd119e938ab054a693cc923b1578437d?p2df

Newton, L. (2019), "Transhumanism Handbook," Springer Publishing, ISBN-13: 978-3030169190, ISBN-10: 3030169197; Kelley, D., "The Sapient and Sentient Intelligence Value Argument and Effects on Regulating Autonomous Artificial Intelligence,"

NESAD; "Collective Intelligence;" National Enviornmental Site Assessment Depository (NESAD), Baltimore, Maryland. Accessed 13 Jan 2022; https://nesad.us/nesad-network/collective-intelligence/

Nørholm, Morten H. H. (2019) "Meta synthetic biology: controlling the evolution of engineered living systems." Microbial Biotechnology 12 (1): 35–37.

Norwood, G.; Deeper Mind 9. Emotions—The Plutchik Model of Emotions; http://www.deepermind.com/02clarty.htm 403 (2/20/02016)

Overgaard, M.; "Measuring Consciousness - Bridging the mind-brain gap;" Hammel Neuro center Research Unit; 2010

Ozkural, E.; "Omega: An Architecture for AI Unification"; arXiv: 1805.12069v1 [cs.AI]; 16 May 2018

Parrots, W, Willcox, G. "Parrotts Classification of Emotions Chart;" http://msaprilshowers.com/emotions/parrotts-classification-of-emotions-chart/ 9/27/2015

Pigott, D.; "Nathaniel Rochester;" 1995; archived from the original on 2011-09-27- http://hopl.murdoch.edu.au/showperson.prx?PeopleID=654

Plutchik, R.; "The emotions: Facts, theories, and a new model." Random House, New York (1962). 26.

Plutchik, R.: "A general psych evolutionary theory of emotion. In R. Plutchik, & H. Kellerman, Emotion: Theory, research, and experience: Vol. 1. Theories of emotion" (pp. 333). Academic Publishers, New York (1980).

Plutchik, R.: "Emotions and Life: Perspectives from Psychology, Biology, and Evolution." American Psychological Association, Washington DC (2002).

Porter III, H.; A Methodology for the Assessment of AI Consciousness; Portland State University Portland Or Proceedings of the 9th Conference on Artificial General Intelligence;

Prince, D.; Interview 2017, Prince Legal LLP

Raven, J., Raven, J.C., & Court, JH (2003, updated 2004) Manual for Raven's Progressive Matrices and Vocabulary Scales. San Antonio, TX: Harcourt Assessment.

Raven, J., & Raven, J. (eds.) (2008) Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics. Unionville, New York: Royal Fireworks Press.

Ray, John. (2012) "The Rosetta Stone and the Rebirth of Ancient Egypt." Wonders of the World (Book 38), Harvard University Press.

Reeves, M.; Whitaker, K.; "The Why of Digital Transformation;" BCG Henderson Institute; 15 Oct 2020; Accessed 16 Oct 2020 @ https://bcghendersoninstitute.com/the-why-of-digital-transformation-34d18f57f9ff

Rescorla, M.; The Computational Theory of Mind; Stanford University 16 Oct 2016; http://plato.stanford.edu/entries/computational-mind/

Samsonovich, A.; "On a roadmap for the BICA Challenge;" June 2012; Biologically Inspired Cognitive Architectures 1:100–107; DOI: 10.1016/j.bica.2012.05.002

Samuel, A.; (1959). "Some Studies in Machine Learning Using the Game of Checkers;" 1959; IBM Journal of Research and Development. 44: 206–226. CiteSeerX 10.1.1.368.2254. DOI:10.1147/rd.441.0206

Sandberg, K; Bibby, B; Timmermans, B; Cleeremans, A.; Overgaard, M.; "Consciousness and Cognition - Measuring Consciousness: Task accuracy and awareness as sigmoid functions of stimulus duration;" Else-vier/ScienceDirect

Schuster, Mike, Melvin Johnson, and Nikhil Thorat. (November 22, 2016) "Zero-Shot Translation with Google's Multilingual Neural Machine Translation System"" in Google Blog.

Searle, J.; "Minds, Brains, and Programs;" Behavioral and Brain Sciences 1980

Seminar Topics; "Measuring Universal Intelligence;" Seminar Report PPT for CSE; Accessed 14 Jan 2022; http://www.seminartopics.co.in/computer%20science/Measuring-Universal-Intelligence.php

Serebriakoff, V, "Self-Scoring IQ Tests," Sterling/London, 1968, 1988, 1996, ISBN 978-0-7607-0164-5

Seth, A.; Theories and measures of consciousness develop together; Elsevier/Science Direct; University of Sussex

Schank, R.; "Conceptual Dependency Theory;" Stanford University; 1969;

Sharkey, N.; "Alan Turing: The experiment that shaped artificial intelligence;" University of Sheffield, BBC 2012; https://www.bbc.com/news/technology-18475646

Silverman, F. (1988), "The 'Monster' Study," Marquette University, J. Fluency Discord. 13, 225-231, http://www.uh.edu/ethicsinscience/Media/Monster%20Study.pdf

Siong, Ch., Brass, M.; Heinze, H.; Haynes, J.; Unconscious Determinants of Free Decisions in the Human Brain; Nature Neuroscience; 13 Apr 2008; http://exploringthemind.com/the-mind/brain-scans-can-reveal-your-decisions-7-seconds-before-you-decideAuthor, F., Author, S., Author, T.: Book title. 2nd edition Publisher, Location (1999).

Solon, O.; "World's Largest Hedge fund to replace managers with artificial intelligence," The Guardian; https://www.theguardian.com/technology/2016/dec/22/bridgewater-associates-aiartificial-intelligence-management

Spark, Andrew (16 Dec 2008). "Oliver Selfridge Computer scientist paving the way for artificial intelligence;" The Guardian. Retrieved 4 Aug 2012.

Spratt, E., et al. (2013), "The Effects of Early Neglect on Cognitive, Language, and Behavioral Functioning in Childhood," Psychology (Irvine). Author manuscript, available in PMC 2013 May 13. Published in final edited form as Psychology (Irvine). 2012 Feb 1, 3(2): 175–182, DOI: 10.4236/psych.2012.32026, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3652241/

Suydam, D.; "Regulating Rapidly Evolving AI Becoming A Necessary Precaution" Huffington Post; http://www.huffingtonpost.ca/david-suydam/artificial-intelligenceregulation_b_12217908.html

Tononi, G.; "Integrated Information Theory;" University of Wisconsin-Madison; Scholarpedia 2015, 10(1):4164; doi:10.4249/Scholarpedia.4164

Tononi, G.; Albantakis, L.; Masafumi, O.; From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0; 8 MAY 14; Computational Biology http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588

Turing, Alan; "Computing Machinery and Intelligence;" October 1965; Mind, LIX (236): 433-460; DOI: 1-.1093/mind/LIX.236.433, ISSN 0026-4423

Urbi, J., Sigalos, M. (2018), "The complicated truth about Sophia the robot – an almost human robot or a PR stunt," CNBC, Accessed May 2019 at https://www.cnbc.com/2018/06/05/hanson-robotics-sophia-the-robot-pr-stunt-artificial-intelligence.html

Vitanyi, Paul; Legg, Shane; Hutter, Marcus (2007). "Algorithmic probability;" Scholarpedia. 2 (8): 2572. Bibcode:2007SchpJ...2.2572H. DOI:10.4249/scholarpedia.2572.

Wang, P.; "AGI Introduction;" Temple University; Accessed 11 Jan 2022; https://cis.temple.edu/~pwang/AGI-Intro.html

Waser, M.; "A Collective Intelligence Research Platform for Cultivating Benevolent "Seed" Artificial Intelligences"; Richmond AI and Blockchain Consultants, Mechanicsville, VA; AAAI Spring Symposia 2019 Stanford

Waser, M.; Kelley, D.; "Architecting a Human-like Emotion-driven Conscious Moral Mind for Value Alignment and AGI Safety;" AAAI Spring Symposia 02018; Stanford University CA;

Watson, J., Rayner, R. (1920), "Conditioned Emotional Reactions," First published in Journal of Experimental Psychology, 3(1), 1-14, https://www.scribd.com/document/250748771/Watson-and-Raynor-1920

Webster (Merriam-WebSter Dictionary) https://www.merriam-webster.com/dictionary/consciousness

Wikipedia Foundation "Moral Agency" 2017 - https://en.wikipedia.org/wiki/Moral_agency

Yampolskiy, R. (2019), "Artificial Intelligence Safety and Security," CRC Press, London/New York, ISBN: 978-0-8153-6982-0

Yampolskiy, R. (2018), "Detecting Qualia in Natural and Artificial Agents," University of Louisville

Yampolskiy, R. (2012), "AI-Complete CAPTCHAs as Zero-Knowledge Proofs of Access to an Artificially Intelligent System," ISRN Artificial Intelligence, Volume 2012, Article ID 271878, http://dx.doi.org/10.5402/2012/271878

Yampolskiy, R.; "Artificial Superintelligence – A Futuristic Approach;" CRC Press; Boca Raton; 2016; ISBN-10: 1138435775