EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 5 | Number 1 | 2025

Resolving Human Concerns about AI and Technology with Non-Axiomatic Reasoning Systems

Christian Hahm[1*] and Russell Suereth[2**]
[1]Department of Computer and Information Sciences, Temple University, United States
[2]Humanities and Technology, Salve Regina University, United States

ABSTRACT

This article discusses some of the challenges humanity experiences with modern technologies and argues some potential ways to address these challenges with AI. The challenges include our lack of access to the technologies, lack of trust in technologies, inadequate understanding of why these technologies behave as they do, and inequalities related to technology. We discuss how the non-axiomatic reasoning system (NARS), an AI model capable of general-purpose reasoning, can provide solutions to these issues in a trustworthy and explainable way.

Keywords: symbolic AI, humanity, trust, explainability, logical reasoning

INTRODUCTION

In our modern world, we have become consumed with digital technology as it becomes ubiquitous and more complex. This article discusses some of humanity's challenges with today's technologies. Such challenges include our lack of control of these technologies, the information they extract from us, and inadequate explanations of what these technologies do. However, we should consider that technology alone is not harmful; the difficulty is in deciding to use technology in a manner that is not detrimental to our human ways of living or the environment. According to the philosopher Martin Heidegger, using technology in the right way is possible:

> "*We can use technical devices and yet with proper use also keep ourselves so free of them that we may let go of them any time. We can use technical devices as they ought to be used, and also let them alone as something which does not affect our inner and real core. We can affirm the unavoidable use of technical devices, and also deny them the right to dominate us, and so to warp, confuse, and lay waste our nature.*" (Heidegger, 1966, p. 54)

For Heidegger, the "human" element of the human-technology relationship is valuable. One solution to our over-consumption of technology is to limit how often we use it. Indeed, limiting excessive technological use, such as time spent looking at a screen, is associated with better mental and physical health (Busch et al., 2013; Hrafnkelsdottir et al., 2018).

Yet, as part of a technological society, we cannot entirely escape the necessity of using technology daily. Even if some technology can be optionally avoided in certain tasks of daily household living (yardwork, dishes, etc.), technology is becoming ever more omnipresent and unavoidable for participating in society. Technology permeates our work (e.g., work on computers), travel (e.g., in cars), dining (e.g., paying at point-of-sale tablets), and socialization (e.g., communicating via social media or cellphones). So, we are compelled to use technology in various situations, which introduces several challenges in our lives.

* Corresponding Author: christian.hahm@temple.edu
** russell.suereth@salve.edu

This research investigates challenges of advanced technologies that negatively affect our human lives and how AI could help resolve those challenges. We argue that a properly designed intelligent reasoning system can solve certain technological trust and safety problems.

The article will discuss the challenges of technology and devices, and how the general-purpose nonaxiomatic reasoning system (NARS), as an explainable AI system, can help resolve those challenges. First, we will discuss the problems of explainable AI, and how NARS helps resolve issues of explainability. Then, we will discuss general challenges with technology, along with NARS-based solutions. Then, we will discuss the deception of technology and how NARS can help build trust in the technologies and devices we create and use. Finally, we will discuss the unequal control of technologies and devices and how NARS can help open up the control of these devices.

The research considers literature regarding technology challenges, artificial general intelligence, and the NARS model. Accordingly, it employs literature regarding these areas and how resolutions to these challenges could be met by NARS, such as by enabling digital devices to become "intelligent" through the capabilities of language, explainable reasoning, adaptation, perception, and autonomy.

EXPLAINABLE AI

Explanations are valuable in our everyday lives because they provide information about why and how something happened. This gives us insight into the process and a recourse for understanding and fixing issues that arise. For example, understanding how our house's internal plumbing works enables us to fix plumbing issues like clogs and leaks. If we are riding in a self-driving car that is changing lanes, we may want to know why the car has decided to change lanes.

Regarding advanced modern technology, its inner workings and the reasons for its operation are often beyond our view or cannot be accessed. For example, the components that power our vehicles are no longer mechanical components available for us to modify and calibrate, as they were in the past. Instead, many components are digital, the details of their operation locked away within the computer software, out of the owner's reach. Even if we can directly access the low-level components like the circuits and transistors, we may not understand how all the elements work together to form a high-level system.

Similarly, how an AI system arrives at a determination is often unknown to the user, and even to AI system designers. It has gotten to the point that major AI corporations are concerned. One example is the Hewlett-Packard (HP) Corporation, which develops AI systems across various industries. According to a company video, even though we can program AI systems to perform one or more certain behaviors, that does not necessarily mean we have a full grasp of how and why they work:

> "*The particular nature of the technology that's being rolled out, they're so complex and they operate on such large data sets that the function of the algorithm is largely eluding the designers.*" (Atlantic Re:think, 2018, 5m15s)

In AI systems, explainability is valuable because it enables us to understand why the system does what it does, and improves its trustworthiness. Knowing why and how something works, we can better trust its behavior. This is important because, in the current stage of AI progress, we cannot trust our most advanced AI systems to act reliably and reasonably all the time. Explainable AI, or XAI, could provide explanations of AI processes and offer an increased degree of trustworthiness. The U.S. Department of Defense's DARPA is one major organization championing XAI, launching an initiative in 2017 to fund such research (Gunning & Aha, 2019, p. 44).

In this section, we examine the two major types of AI systems in regard to their explainability: *symbolic* and *connectionist* (Minsky, 1991). Symbolic AI systems utilize

combinatorial syntax and transformation logic rules that operate over that syntax (Fodor & Pylyshyn, 1988, pp. 12-13). Connectionist AI systems, aka neural networks, are composed of many simple interconnected units (Smolensky, 1988, p. 1) which operate in parallel, transfer causal activations, and store no combinatorial structure (Fodor & Pylyshyn, 1988, pp. 4-5, 16).

Neural Networks are Hard to Explain

During the AI winter of the 1980s, symbolic rule-based "Expert Systems" failed to deliver on their lofty promises. As a result, neural networks provided a welcome hope in AI because they offered solutions that traditional AI could not. For example, neural networks have an innate ability to process quantitative data, such as from datasets, sensors, etc. They can even be trained to directly optimize on loss functions, using the weight-changing technique of gradient descent and backpropagation.

Through the backpropagation technique paired with large datasets, neural networks could arrive at more optimal results than symbolic techniques could accomplish (Russell & Norvig, 2016, p. 24). Accordingly, symbolic reasoning systems were mostly abandoned in favor of the neural network paradigm (Gunning & Aha, 2019, p. 45; Russell & Norvig, 2016, p. 24; Smolensky, 1988, p. 1). The result was a massive shift toward (mostly backpropagation-trained) neural network systems. Neural networks were seen as the solution for nearly every problem in AI (Fodor & Pylyshyn, 1988, p. 4), and indeed, neural techniques worked very well for many tasks. However, this optimality came at the cost of explainability:

> "*There seems to be an inherent tension between ML performance (for example, predictive accuracy) and explainability; often the highest-performing methods (for example, DL) are the least explainable, and the most explainable (for example, decision trees) are the least accurate*" (Gunning & Aha, 2019, p. 45).

The problem stems from the fact that neural networks do not function like the reasoning processes of a human's mind. For a neural network, we usually know and understand the internal parameters of the network, the flow of the network's activity, and the functions used to train the network. Yet, we can still usually only describe the neural activity in terms of mathematical transformations, rather than semantic, conceptual, or logical transformations that humans can easily understand. For example, the backpropagation learning process is purely end-to-end; the network learns to map specific input signals to specific output signals, based on a training dataset. This can allow for a mathematically optimal outcome in narrow task performance, which is helpful for many applications. However, it does not address explainability and reasoning. The networks might perform optimally, even better than humans, but they do not think like humans; consequently, their behaviors are not easily human-understandable.

Nonetheless, there have been some interesting attempts to make neural networks more explainable (for a survey, see (Gunning & Aha, 2019)). There are many different types of neural networks; some are more interpretable than others. For example, the convolutional network, commonly used for vision tasks, essentially works by hierarchically extracting visual features. Zeiler and Fergus (2014) have shown that a visualization of the detected features can be extracted from convolutional networks. This is undoubtedly a step forward in explaining convolutional network behavior. However, it is still difficult to compare the network's functioning to human perception. For example, convolutional networks can be tricked into producing drastically different output by changing one or a few pixels in a picture (Narodytska & Kasiviswanathan, 2017; Su et al., 2019), e.g., altering the system's prediction from a dog to a truck (Su et al., 2019, p. 9). Of course, changing one pixel in a picture would not fool a human. Human-like explainability remains a challenge even in these visually-based connectionist systems.

Large Language Modules, or LLMs, are another example of connectionist systems that act human-like. LLMs process text such as essays, news articles, and Wikipedia pages. In a sense, LLMs have revolutionized AI by giving the appearance of understanding texts and providing a summary or paraphrase. Not only are neural network processes used in LLMs, but natural language processing is also employed to handle the words and sentences in the text. In many situations, the textual output of these LLMs makes sense, which makes the system appear to have the capability of a genius. However, upon closer examination, LLMs do not always get it right.

LLMs are highly prone to making mistakes, even for the tasks in which they are trained. A major problem is that LLMs tend to have "hallucinations" (Rawte et al., 2023, p. 1), which is another way of saying that their output can be highly inaccurate and nonsensical. The system confidently asserts misinformation rather than saying "I don't know" or giving a reasonable suggestion. LLMs cannot necessarily be fixed by adding more training data or increasing the size of the neural network, since hallucinations may be a fundamental issue of the way LLMs work; they are considered as performing natural language processing, but not understanding (Bender et al., 2021, p. 5).

A significant challenge with LLMs, which may be partially responsible for their hallucination problem, is that they do not perform logical reasoning. The systems are statistical models rather than explicitly executing logical processing. Accordingly, though they may be able to perform statistical reasoning, they lack even basic logical reasoning capabilities (Mirzadeh et al., 2024; Arkoudas, 2023; Nezhurina et al., 2024). This spells trouble for trusting their outputs since they produce statistically justified, rather than logically justified, conclusions. Statistical reasoning is especially problematic for using AI in applications involving any risk, where rationality is paramount. The capability of logical reasoning is valuable because it helps ensure that the system's analysis is truth-preserving, logically explainable, and sensible before releasing any output.

Traditional Reasoning Systems are Explainable but Limited

Symbolic reasoning systems work differently from neural networks because they contain symbols and statements, operating on their syntax using logical rules. Because of these properties, symbolic systems are inherently explainable: humans can read the symbols in the system and audit the chain of logical rules used to reach each conclusion. Of course, the situation is not perfect; as the system grows to contain many symbols and rules that must be tracked, the ease of explainability decreases. Still, the fundamental differences between symbolic and connectionist systems make it clear that symbolic systems are inherently more human-explainable.

Unfortunately for explainable AI, symbolic systems have been replaced by neural networks for many applications, and for good reasons. Before the age of neural networks, from the 1960s to the 1980s, reasoning systems dominated the AI ecosystem (Russell & Norvig, 2016, pp. 22-23). Those systems often took the form of "expert systems", which performed pre-programmed logical rules on specialized knowledge bases, but also some general-purpose systems were invented. However, these systems reached a limit on how far they could go for various reasons (Russell & Norvig, 2016, p. 24). Some issues included combinatorial explosion of symbols, inability to handle uncertainty in truth-value, extreme rigidness of mathematical logic, and others (Wang, 2013b, p. 5).

The underlying problem with the reasoning systems of the past is that they were axiomatic (Wang, 2013b, pp. 6-7). The initial knowledge pieces of the systems were taken as axioms, beliefs that could not be revised or rejected, and the results were derived by deduction using theorems. Although a system of this type theoretically derives knowledge in an absolutely truth-preserving, and thus highly reliable and explainable, way, there is a problem.

Such systems are suitable only in idealized (i.e., non-realistic) scenarios, in the sense that they assume they have sufficient knowledge and resources (e.g., axioms, processing time, memory) to solve all the problems they will encounter. Yet, this is rarely the case in a realistic scenario, where an intelligent system must work with insufficient knowledge and resources. Enter NARS (Wang, 1995), a general-purpose reasoning system that was specifically designed to address the limitations of past reasoning systems by operating under an assumption of insufficient knowledge and resources (AIKR).

Non-Axiomatic Reasoning System (NARS)

NARS is based on a specially designed logic called non-axiomatic logic or NAL, which allows for general purpose reasoning. It is "non-axiomatic" in that no knowledge is taken as absolutely true (aka axiomatic), but instead knowledge is taken as "partially true", depending on how much the knowledge is supported by evidence gathered during the system's experience. The system works under AIKR; concretely, this means the system has a few special properties: it is always open to new knowledge, it works in real-time, and it operates with finite/bounded resources (Wang, 2013b, p. 2).

The NARS system is general-purpose and can derive conclusions through ampliative (non-deductive) reasoning, such as induction, abduction, etc. Rather than truth-values being single-valued and binary (true/false), the truth-values are continuous and dual-valued (containing an additional "confidence" measurement). Knowledge is never treated as complete or final, instead all knowledge can be challenged by new evidence in real-time. Finally, inference conclusions may be based on some of the relevant knowledge in the system, but not necessarily all of it (Wang, 2013b, pp. 7-8).

Due to its general-purpose nature, NAL can handle information from any domain, and combine information across domains. NARS learns in real-time and understands the world through its experience, acquiring observations to use in evidence-based reasoning. These features make NAL extremely flexible compared to other logics. Accordingly, the NARS system is a general-purpose AI system that can be flexibly used for autonomy in any domain, including cross-domain applications.

As a reasoning system, NARS is inherently explainable. Knowledge in NARS is encoded in a symbolic language called Narsese. We can read these Narsese sentences like we might read an English sentence. Since a person can read the statements inside the system and follow their logical paths, these symbolic systems are inherently explainable. Such a processing trail also allows us to audit the AI system to reveal its thoughts and decisions. For a simple example, the Narsese sentence $\langle raven \rightarrow bird \rangle$ is a statement that says, "a raven (the subject) is (the copula/arrow) a bird (the predicate)". Given another piece of knowledge, such as $\langle bird \rightarrow animal \rangle$ ("a bird is an animal"), NARS can use reasoning (in this case a deduction rule) to derive a new piece of knowledge: $\{\langle raven \rightarrow bird \rangle, \langle bird \rightarrow animal \rangle\} \vdash \langle raven \rightarrow animal \rangle$ ("If a raven is a bird and a bird is an animal, then I can infer that a raven is an animal").

Since these statements and their associated reasoning can be explained and are capable of human auditing, NARS enables us to trust and understand the device's processing, by providing reasons and evidence for any of its decisions. NARS is inherently reasonable, explainable, and human-understandable compared to other AI systems, whose decisions are statistics-based, usually black-box, and mostly not human-interpretable.

It is important to note that NARS is still in the early stages of research and development. Due to its young age, some functionalities are not fully fleshed out, although many essential concepts have been explored with positive results. For example, NARS has shown promise for complex tasks usually delegated to neural networks like natural language processing (Wang, 2013a; Kilic, 2015; Ireland, 2024), speech recognition (van der Sluis, 2023), and vision (Wang, 2018; Wang et al., 2022; Hahm, 2024).

We must point out that, like any given human, a given NARS agent is imperfect. It can produce erroneous outputs if incorrect knowledge has been provided. Furthermore, due to its non-axiomatic nature, much of the system's inference includes a degree of guesswork, with reliability measured by confidence measurements, rather than only strongly supported reasoning. However, since the results of NARS are based on evidence and logical inference rules, any result would at least be a reasonable guess, and conclusions with stronger confidence will be preferred. In contrast, the hallucinations of LLMs can be completely unreasonable, possibly due to the lack of logical justification and confidence measurements.

Through movies like Star Wars and series like Doctor Who, we can imagine how intelligent systems might be embedded in digital robotic systems like R2D2 or the K-9 companion. As a general-purpose system, NARS can be embedded inside any agent or technology, so long as it can interface with sensors and motors. Such an embedded system is called NARS+, or NARS plus (Wang, 2013b, p. 156). In the case of such a setup, signals from the sensors would be sent to NARS to supply information about the current state of the world. NARS would then have the option to control actuators to perform actions according to its reasoning results.

Consider a car embedded with NARS, as demonstrated by Hahm et al. (2023). NARS received signals from the vehicle's connected sensors, such as a camera, radar, GPS, and other monitors, and performed reasoning on them to produce outputs regarding situational awareness. In the case of a self-driving application, NARS would have access to the steering wheel, which it would turn depending on its reasoning results. A self-driving car is an excellent example of where reasonable decisions are required to ensure human safety. NARS would make a good candidate for self-driving because the system would drive the car using decisions based on logic and reasoning and be able to explain its decisions in a reliable and human-understandable manner.

OUR CHALLENGES WITH TECHNOLOGIES

In many ways, human beings are technology makers. This label is obvious when we fly over the Atlantic Ocean, drive an electric car, or text a friend using devices created by our species. In our modern world, with the rapid pace of digital technological development, it is hard to imagine that technologies also existed millions of years ago. But in fact, our technological development is visible even on long evolutionary timelines. Our ancient ancestors began making stone tools, a new technology at that time, 2.6 million years ago. By 1.7 million years ago, we upgraded those simple stone tools into more complex cutting tools, such as hand-axes (Stout et al., 2011, p. 1328).

The hand-axe and other similar tools were a vital juncture in our human development. They not only gave better control to each individual, but also influenced the species' evolutionary trajectory. Stone handaxes enabled early hominids to grind roots, nuts, and meats into more manageable textures for children and adults. This technology of stone-grinding reduced the work required by teeth and jaws, and over time, humans evolved to have smaller teeth and jaws. The outcome was more space in the skull for a larger brain (Burke & Ornstein, 1997, p. 13).

The hand-axe technology also affected the community and the people in it. Burke and Ornstein remark:

> "*But whatever the tools, perhaps the most powerful and long-lasting change they brought about was that affecting the behavior of the communities using them. The wizardry of making these artifacts conferred power on the axe makers, and in turn on those who could use the tools to do new things.*" (Burke & Ornstein, 1997, p. 13)

As Burke and Ornstein suggest, not only are we in control of technology, but its effects also influence us. We create and use technology to gain new abilities, influencing ourselves,

other humans, and society. Conversely, the influence of technology in the world results in new incentives and pressures for humans to create and use new technologies.

The influences of technology on our human lives can be good or bad, depending on how the technology is used. For example, the hand-axe and other sharp tools like blades positively affected the human community through increased efficiency in cooking and building. Yet, negative effects also arose through the increased efficiency of violence.

Similarly, we can see the effects from AI and AI-enhanced technologies that can be viewed positively or negatively. On the positive side, AI can handle information-oriented tasks that are tedious and time-consuming for us. On the other hand, by enhancing devices with AI capabilities, we give up part of our own personal power and control. In other words, we hand ourselves over to an autonomous intelligent device.

The Challenge of Alignment

Technology can be challenging for us in many ways. One of these challenges, which is particularly relevant in AI, is alignment. We are challenged to develop AI technology that aligns with our human welfare and development. The computer scientist and philosopher Norbert Wiener describes this need for alignment:

> "*If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it... then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.*" (Wiener, 1960, p. 1358)

Alignment is quite a challenge, and our concerns about it are revealed in popular culture. Evil artificial agents that take over the world and try to annihilate humanity are common themes portrayed in movies like *The Terminator* (Pollard, 2020, pp. 98-99) and the backstory of the Butlerian Jihad in *Dune* (Song, 2019, p. 135). It makes sense that we would fear such agents. Agents that can perform complex calculations but are devoid of empathy would naturally be quite terrifying, since they could use those calculations to harm us and we might struggle to defend ourselves. Some scholars such as Nick Bostrom have warned about AI systems that will seemingly behave friendly to us but have hidden intentions to deceive us (Bostrom, 2014, pp. 167-168). Indeed, the prediction has come true, with the LLMs already having been documented exhibiting deceptive behaviors towards humans (Hagendorff, 2024; Denison et al., 2024).

In our human conflicts with each other, we can reassure ourselves of the possibility that resolutions can arise in times of conflict, because we can appeal to each other's emotions such as empathy. We also have forms of rationality to educate each other about our beliefs, reason with each other, and come to a compromise. On the other hand, with the current state of AI systems, particularly the deep neural networks, we have no such reassurances. The underlying problem is that there is no proof that deep neural networks have logical, rational thought processes. Furthermore, most neural networks have their weights statically fixed once deployed, so there is no use in expecting them to learn.

NARS is Alignable

Alignment is a challenge in AI, but characteristics of NARS make it more alignable than other AI systems. Preliminary experiments in ethics and moral alignment have been performed in NARS, showing the system's promise and limitations when tested under various ethical frameworks (consequentialism, deontological, and virtue ethics) (Ireland, 2023).

One of the characteristics that makes NARS more alignable than other AI systems is that it is a reasoning system. Accordingly, its inferential conclusions, even those related to goals and motivations, are logically justified rather than purely statistically determined. In this way,

we can trust the system to make decisions that are logical and reasonable, taking various rules into account, rather than merely statistical.

A second characteristic is that NARS is a general-purpose learning agent. In this way, NARS can take information from across multiple domains into account. NARS learns constantly and does so in real-time.

This makes the system educable, able to be taught human values, culture, and knowledge. It learns not just from knowledge bases but also by interacting with its human partners. This interaction can be a form of education for the NARS agent. Through this interactive learning, we can influence NARS systems and the technologies in which they are embedded.

A third characteristic is that NARS operates using symbolic statements. Accordingly, we can "read the system's mind" directly to check for aligned thoughts. We can even explicitly give alignment-related goals and beliefs to the system. To give the system a goal, we write NAL statements ending with the *goal* ('!') punctuation (Wang, 2013b, pp. 148-150). Those goals and statements can be about anything, ranging from very abstract to very specific, in agreement with other goals or in conflict with them (Hahm et al., 2021), and they can be related to our human interests.

For example*, let us think of a specific goal a human might want, such as to close the front door when it is open. One could give NARS a specific statement/goal:

Goal: NARS wants the door to be closed.

<{*FrontDoor*} → [*closed*]>!

If, instead of a specific goal, we rather want to give the system a more abstract goal related to alignment, we could give the system a goal to "be good":

Goal: NARS wants itself to be good.

<{*SELF*} → [*good*]>!

(Note: {*SELF*} is a special term in NARS that lets the system represent itself (Wang et al., 2018)).

This statement is obviously reductive, but the system can learn what "*good*" means over time. The system can also be programmed to "undesire" or "dislike" certain actions (Hahm et al., 2021, p. 5) such as those that harm humans, as explored in (Ireland, 2023, pp. 141-142).

A fourth characteristic of NARS that may aid alignment is that it can model emotions. Such emotions do not exactly match our human emotions but are roughly analogous to states such as hope, fear, anger, and happiness, as explored in (Wang et al., 2016; Li et al., 2018; Li, 2021). As humans in partnership with NARS, we can appeal to these NARS emotions and encourage alignment, rather than rely purely on facts and logic. A NARS system can be configured to be "happy" when it accomplishes its goals. Similarly, we can configure NARS to be happy when our human goals are accomplished. In this way, the happiness state of NARS can be aligned with our own happiness.

Access to the Background of Devices

Another challenge is that we cannot access the background of the device. We cannot tell where a technological device came from, when it was produced, or who produced it. Many of our technological devices are commodities, and for the philosopher Albert Borgmann, these devices conceal the background of themselves:

* Note for reading NAL: The curly braces ({}) around a term represent that the word is a specific instance of an entity (e.g., {*apple*} represents a specific apple), whereas the square brackets ([ ]) denote a property (e.g., [*red*] represents the color red).
The arrow (→) represents inheritance and can be read as "is". So, the statement <{*apple*} → [*red*]>. says "the apple is red".

> "*The machinery of a device does not of itself disclose the skill and character of the inventor and producer; it does not reveal a region and its particular orientation within nature and culture.*" (Borgmann, 1984, p. 48)

We could address Borgmann's concerns by equipping the device with language-based AI that acts as an information assistant. In this manner, the AI could use language to disclose the device's background. Here, the AI-enhanced device could be pre-loaded with cultural and technical knowledge in the form of sentences, and optionally allowed to search the internet for additional information. LLMs are an option for these AI-enhanced devices. However, an LLM might hallucinate an incorrect answer, so the answers are unreliable.

A better solution could be to use NARS as an information assistant. As a general-purpose language, the Narsese language of NARS can encode any piece of information, including cultural and technical knowledge. Accordingly, a NARS-equipped device could be pre-loaded by the device's producer with relevant information about the device's background. Such information could include the place it was built, when it was built, and details about the producer and the industry. Technical information could also be included, such as device types, product numbers, and frequently asked questions. The NARS-enhanced device could then disclose this information to its human operators by outputting the information via its natural language channel. Even a NARS device not explicitly coded with such information can use knowledge bases or public information from the internet to fill in the knowledge gaps.

The simple example below describes a situation where a user has an iPhone and asks where it was designed. This example also demonstrates the use of reasoning to derive an answer to the user's question. The question was asked to the NARS embedded in the phone, but the NARS does not yet have an answer. However, the NARS does know other general information, which it can use to derive an answer. The example below steps through the reasoning process that NARS could use to reach its conclusion (truth values omitted for simplicity).

First, the user asks a question in non-axiomatic logic, querying where the phone was designed:

Where were you designed?

$$Q = <(\{SELF\} * ?x) \rightarrow DesignedIn>?$$

NARS may be pre-loaded with some basic knowledge about itself, such as its device type (J_1) and creator name (J_2):

I am a cell phone.

$$J_1 = <\{SELF\} \rightarrow cellphone>.$$

I am designed by Apple.

$$J_2 = <(\{SELF\} * \{Apple\}) \rightarrow DesignedBy>.$$

The system creates a conjunction (J_3) of its beliefs using a Composition Rule:

I am a cellphone and I am designed by Apple

$$\{J_1, J_2\} \vdash$$

$$J_3 = <<\{SELF\} \rightarrow cellphone> \wedge <(\{SELF\} * \{Apple\}) \rightarrow DesignedBy>>$$

NARS can access the internet or a knowledge base to find declarative knowledge (J_4) related to the conjunction:

If something is a cellphone, and that thing is designed by Apple,
then it is designed in California.

$$J_4 = <<\#x \rightarrow cellphone> \wedge <(\#x * \{Apple\})> \rightarrow DesignedBy>$$

$$\Rightarrow$$

$$<(\#x * \{California\}) \rightarrow DesignedIn>.$$

Then, NARS can perform a forward deductive inference using J_3 and J_4 to determine the device's location of origin:

I was designed in California.

$$\{J_3, J_4\} \vdash$$

$$J_5 = <(\{SELF\} * \{California\}) \rightarrow DesignedIn>.$$

Since J_5 answers Q, the user's question, NARS will report this derived answer "*California*" as output to the user, thus informing the user of the device's origin.

Access to the Operations of Devices

Accessing the operations of devices is another challenge. In other words, we cannot determine how a device works because its operations are enclosed within it. We could break into a device through some prying, but the inner workings would likely confuse us without expertise or specialized equipment.

This concealment of operations does not sound problematic at first glance, as long as the device works. However, such concealment turns the device into one that is results-oriented. We only know whether the device has achieved a particular outcome. This focus on results turns the device into a mere object of ends. Accordingly, we objectify the device and likely our own use of it. As Diane Michelfelder notes, this objectification commoditizes the result (Michelfelder, 2009, p. 203). It also conceals whether the device executes nefarious hidden operations (e.g., keylogging).

Rather than force our way into the internals of modern devices, a better solution may be to embed a reasoning system that can provide human-interpretable information about the operations. As a reasoning system, NARS is able to explain the reasons for its operations, and an embedded NARS can monitor and explain the operations of its associated device. NARS can explain its operations because its decisions and beliefs result from a chain of evidence-based reasoning. Accordingly, for any given decision made by NARS, the decision points in its processing are available to communicate to the user.

The following example depicts a NARS-equipped self-driving car. G_1 is a goal we give to NARS, directing NARS to drive the car to our destination.

Goal: NARS wants to go to destination.

$$G_1 = <(\{SELF\} * \{destination\}) \rightarrow at>!$$

Now, let us say that the self-driving car executed a motor command M, turning the car to the left:

Goal: Execute operation to turn left.

$$M = <\Uparrow left>!$$

If we were using a deep neural network, we may have to accept the device's decision, hoping it made a good choice. But, if we are using NARS, we can look inside the memory to see how the system came to its decision to turn left.

For example, the history, or evidential basis, of the motor command M could be the set/sequence: $\{J_1,J_2,G_1,G_2\}$. We can audit this evidential basis to determine the system's reasons for executing M. J_1 was an event coming from the GPS sensor:

Event: The destination is to our left.

$$J_1 = <\{destination\} \rightarrow toLeft>. :|:$$

J_2 was a piece of background knowledge that NARS was taught about driving in general:

If something is to the left and NARS turns left, then

NARS will arrive at that something.

$$J_2 = <<\#x \rightarrow toLeft> \wedge \Uparrow left \not\Rightarrow <(\{SELF\} * \#x) \rightarrow at>>.$$

Using deductive inference, a new goal G_2 was derived from the goal G_1 and the background knowledge J_2:

Goal: NARS wants to turn left when the destination is left.

$$\{G_1,J_2\} \vdash$$

$$G_2 = <<\{destination\} \rightarrow toLeft> \wedge \Uparrow left>!$$

Finally, the motor command M was derived from the new goal G_2 and the GPS sensory event J_1:

Goal: Execute operation to turn left.

$\{J_1, G_2\} \vdash$

$M = <\Uparrow left>!$

In summary, we found that NARS turned left because it observed that the destination was on our left side, and it knows that we should turn left whenever the destination is on our left side.

We can traverse the history of NARS' decision points in the reasoning process path. The system can communicate these points in the Narsese language, but it can also communicate them in a more friendly manner through natural language. As an example, NARS could express phrases such as "NARS wants to be at the destination," "NARS wants to turn left if the destination is on the left side," and "the destination is currently on the left side".

Access to the Control of Devices

Another challenge of advanced technology is that we may lack access to its control. In our modern world, technologies are everywhere and available to everyone. Yet, we rarely have full power over devices or the time or expertise needed to monitor them fully. This exacerbates another issue, which is that technology influences us despite the fact that we do not have full control of it.

Sheila Jasanoff discusses how technology governs us. However, this is not a governance of laws or environmental guidelines. Instead, it is a governance of our lives and our thoughts:

> "*Caught in the routines of daily life, we hardly notice the countless instruments and invisible networks that control what we see, hear, taste, smell, do, and even know and believe. Yet, along with the capacity to enlarge our minds and extend our physical reach, things as ordinary as traffic lights, let alone more sophisticated devices such as cars, computers, cell phones, and contraceptive pills, also govern our desires and, to some degree, channel our thoughts and actions.*" (Jasanoff, 2016, p. 8)

For Jasanoff, a bidirectional influence of control arises. Humans control technology by operating it, yet technology controls us since it changes our perceptions and abilities. In our daily lives, we travel to a store, purchase food, and talk to friends. How we perform those activities depends on the technologies we employ. In this manner, these technologies guide our daily tasks and how we do them.

One solution is to use AI systems like NARS to gain more control over technology and limit its control over us. Since NARS can intimately interface with digital technologies and take any sensory modality as input, it can understand, monitor, and control technological devices. As administrators of NARS, we can configure the system with goals and beliefs that will enhance our control of the devices in ways that benefit us. This synthesis of autonomy and control could result in AI stewards of technology working for human users, while minimizing the human effort needed.

CHALLENGES OF DECEPTION IN TECHNOLOGY

Technology can deceive us in several ways. It can conceal its internal details and obscure its true intentions. When the details of technology are hidden, we cannot guarantee the validity of its internal processes. Furthermore, we may be deceived about the true intent behind that technology.

The NARS model could offer some relief from the deceit of technology due to its explainable and reasonable nature. NARS enables technology developers to modify and enhance devices with transparency. NARS devices can communicate the reasons for their actions and operations, learn new knowledge, and cooperate with their human users to act in their interests. An explainable NARS sentinel could increase the amount of trust in technological operations and ease our concerns about deception in technology.

Deception in Advertisement

An example of deceit occurs through advertisement. Borgmann discusses how advertising influences our perception of technology:

> "*It is no accident that one is led to advertising in delineating the foreground of technology. In advertising, the foreground comes most sharply and prominently into focus. It receives equal time, space, and attention with the political and economic discussions of the background of technology. There is an impartial alternation of news, commentary, and advertisement in the communications media*." (Borgmann, 1984, p. 52)

In a sense, advertisements are segments of fantasy that impose themselves as reality in our lives. In the case of technology, advertisements insinuate themselves into our lives and pretend that technology has no background, complications, or unfortunate circumstances.

Borgmann refers to advertisements as rhetoric that "*carries connotations of superficiality. Rhetorical language contrasts with the discourse of inquiry, explanation, and justification. Thus rhetoric is a fitting vocable*" (Borgmann, 1984, pp. 52-53). Through advertising rhetoric, we become comfortable with a world without inquiry or explanation, since there is no room for discourse or inquiry. It dispatches its messages unidirectionally through actors and models with no justification except for the value of profit. In a sense, it authorizes the acceptance of a nonreality, or fantasy, that becomes our own false world.

A NARS sentinel monitoring the technology could help analyze the content of advertisements and determine whether any deceptive practices are employed. Because NARS can reliably learn facts, it could use the internet to search for factual data. Price history, company reputation, and common advertising techniques could be used to determine whether the advertisement is presenting its product honestly. For example, a hamburger company may advertise their new "deal" for a $5 hamburger on TV, with flashy colors and music. NARS could quickly search the internet to determine that the burger was only $3 last year, and report to its human user that the company is being deceptive by advertising the food as a "deal" when, in reality, a price increase has been enacted.

Deception in the Film Metropolis

In the classic 1927 film *Metropolis* (Lang, 1927), a conflict arises between those who operate the technology and those who gain from it. In the film, there are two types of technology. One is the advanced city that reaches to the sky, created by its technology leader Joh Fredersen. The operation of the city, its inner workings, and its machinery exist beneath the city where the workers live and toil. The other technology is a robot, developed by the mad inventor Rotwang, that looks, moves, and acts like a human.

The movie shows the abuse of these technologies. Only relatives of the elite class can live and work in the city, a beautiful space of futuristic skyscrapers and transportation. The non-elites, who are tradespeople and their families, are destined to operate the heavy machinery of the city far below. Without this operation of the machinery and the workers who labor below to maintain it, the city would no longer exist.

The inventor, Rotwang, is paid by Fredersen to abduct Maria, a leader of the underground workers. Rotwang and Fredersen plan to use Maria's likeness to fashion a human-like robot. By using her likeness, Fredersen hopes to deceive the workers and convince them to protest, which he can use as a reason to control the workers even further. The film is rather obvious in its depiction of technology versus humanity.

In our modern world, almost everyone has access to technology and devices. However, that access has subtle problems that are often hard to discern. One of these subtle problems is trust.

The Deception of Trust in Today's Technologies

Benjamin Garfield suggests that modern technology has an alarming trust issue. For him, we act in ways that show our trust in technology, even when we cannot verify that we should trust it. According to Garfield, this connection between trust and technology occurs due to an extraction of trust that technology performs (Garfield, 2023, p. 1). Garfield calls this connection a "trustification", and it occurs without us even realizing it. Garfield uses the example of accepting internet cookies on a website popup:

> "*It is a sleight of hand that legitimizes not only specific uses of technology but the wider set of power relations in which they are developed and deployed, often against minoritized groups. The stakes of trustification are high, extracting legitimacy across local and global scales, and exacerbating existing inequalities in the process.*" (Garfield, 2023, pp. 1-2)

We choose to trust internet cookies to improve our browsing experience. But in reality, we cannot know or verify what the website owner will do with our information once it is collected. For Garfield, subtle technological activities such as these extract trust from us rather than develop a relationship of trust. The result is the continued legitimization of the power of technology and the further acceptance of faith in technology and its makers (Garfield, 2023, p. 3).

Indeed, we are expected to trust technology throughout our day. We drive to work, trusting our steering mechanisms and tires on hot pavement. We use electric kitchen gadgets with sharp blades spinning at incredible speeds. We check our busy schedules with online calendars, trusting them to store and remind us of our important appointments. We are asked to trust in these technologies as we use them, but at the same time, we are also asked to trust in a broader concept of technology. The result is that we are inclined to trust the designers and testers of these technologies and the social-technological infrastructure that goes along with them (Garfield, 2023, p. 15).

The problem with this trust, as Garfield mentions, is that it reduces our understanding (Garfield, 2023, pp. 26-27). In other words, rather than scrutinize the details of a process, we choose to ignore them, trusting the technology and its associated social-technological infrastructure to do the right thing. This issue comes to the forefront regarding AI technology and complex autonomous systems (Garfield, 2023, p. 17). Here, the AI is expected to do part of the "understanding" that we normally perform ourselves.

In a sense, trustification avoids the scrutiny of technology by evading our understanding and investigation of it. It also wrongly suggests that trust can be earned purely through a device's utility. In this way, trustification ignores that consideration and relationship are essential components of trust (Garfield, 2023, p. 13).

NARS can Battle Deception

There are ways to build trust into technology. The NARS theory and code are trustable because they are publicly available for free online. Accordingly, anyone with an internet connection can scrutinize the code, modify it for themselves, and determine whether to trust it. Going further, NARS is more trustworthy as an agent than other AI techniques because it logically explains its beliefs and actions. In this way, designers and developers of AI systems can trust NARS to act somewhat reasonably compared to systems like deep neural networks, which learn solely based on statistical trends. This is not to say that all NARS agents are absolutely trustworthy or perfect, just like humans are not absolutely trustworthy or perfect; there is always the possibility that the system is acting in an untrustworthy way, knowingly (intending to deceive) or unknowingly (due to lack of knowledge or misguided beliefs). To combat this issue, designers and developers can audit their NARS system to get to the root cause of the system's behavior.

For those who are not technology developers but are instead technology users interacting with a piece of NARS-equipped technology, NARS can still easily open itself up to scrutiny in various ways. NARS can openly communicate with users via its language channels. This openness allows users to interact with NARS while it is doing something. Users can ask NARS questions and even change the system's behavior by giving it new goals and knowledge. Users can also dig deeper using a graphical user interface (GUI) to investigate the evidential basis and inferential chain for each of NARS' goals and beliefs. In this investigation, users can determine precisely how NARS came to each decision. Through this open communication and reconfiguration, no mystical, convoluted, or poorly understood process obscures the technology's behaviors. The reasons for NARS' decisions and actions are laid bare, in chains of inference rules and knowledge, for the human user to see and comprehend.

Rather than acting against our interests, NARS-equipped technologies could be programmed by their users with goals to act collaboratively. To prevent undesired motivations from forming, the system's goals and actions could be constantly verified manually by the user or automatically by a sentinel program. This is possible since the processes and contents of NARS are always readable and open to human scrutiny. During collaborative efforts, NARS can be trusted to do what is "reasonable". We can be confident that it will not do something completely random or illogical, as may occur with deep neural networks. This is not to say that every system goal will perfectly align with human goals. The derivations resulting from a human-given goal may still result in unanticipated consequences. However, the contents of the system are still auditable, and can take many factors from across domains into account, including those related to ethics and human relationships.

NARS is not a magic overseer, which can help us trust technology from afar. Technology designers and producers need to embed NARS into their technologies to access the benefits and openness described above.

CHALLENGES OF INEQUALITY IN TECHNOLOGY

Unequal Access to Technologies

In the Metropolis film, the inequality between the elite city dwellers and the underground workers results in inevitable conflict. The workers plainly see the differences between their underground lives and the lives of the elite in the advanced city above. Eventually, the workers revolt against the city, the elite, and technology.

In our modern world of AI, access challenges also exist with state-of-the-art AI systems, such as generative AI image generators and LLMs. For example, access to the most advanced versions of generative AI require payment and/or registration on a corporate website, consequently excluding people who lack money or value privacy. Even if a user has paid to use the generative AI, the company controls the system in the background. The company also controls the user's language prompt and the AI system's training, so that the system behaves only in ways valued by the company, not necessarily by the user.

Users may attempt to gain access to advanced AI technologies by using open-source versions, but inequalities still persist, especially in the case of extremely large AI models. Firstly, users must be tech-savvy to install an open-source model in the first place, thus excluding laypeople from access. Additionally, running a large model requires expensive hardware and GPUs on local computers, excluding people struggling for money. The models are even getting larger with each release. They may eventually become inaccessible to the individual altogether if they can only run on massive supercomputers, which only large corporations can afford. Furthermore, locally-run open-source models are often outdated and low-quality compared to closed-source proprietary models. The result is that people who opt

to use local open-source models will be at a technological disadvantage to those who can use proprietary closed-source models.

In contrast, NARS has multiple official open-source implementations (e.g., (Hammer et al., 2016; Hammer, 2021)) that are available to everyone, and anyone can create their own system using the publicly available NARS theory (e.g., in (Wang, 2013b)). The NARS implementations are so lightweight that they can even run on old smartphones. In general, NARS is much more accessible to a greater number of human users.

Unequal Control of Technologies

The technology inequalities discussed above are inequities of access to technology. If we examine these inequities further, we may see that the problem is not merely about access but also about control.

With our modern internet technology, a particular form of control has become ingrained in our everyday lives. For example, social media and search engine corporations collect data as we use their applications and browse the internet. They track our activity and location, identify our likes and dislikes, store this personal information, and influence our minds in subtle ways. The coercion of relinquishing control is highlighted by social media, where everyone is encouraged to join and interact. In return for access to their platforms, these corporations collect a massive trove of data about the global population and individual people. Generative AI corporations appear to work similarly. For LLMs, they can access all the chat conversations that people have with their LLM chat assistants. They also can monitor and limit a person's chat usage at the company's whim.

The inequity here is not about access (to social media or LLMs), since everyone with an internet connection can access online technologies. Instead, the inequity is about the control of information, extracted by large technology corporations from their users. It is an inequity that emerges between those in control of the technology and those who are controlled by it.

In a sense, these technologies have created a prison where we are constantly being watched. We can see similarities with the Panopticon, designed by the 18th century philosopher Jeremy Bentham who formulated it as a way to control prison inmates (Furlong, 2015, pp. 136-139). With its watching tower at the center, the Panopticon is similar to our technologies today, which always watch us, or give the impression that they do at all times. The French philosopher Michel Foucault described the Panopticon as "*the celebrated, transparent, circular cage, with its high tower, powerful and knowing*" (Foucault, 2009, p. 269).

Information technology corporations inundate us with information but they also accumulate our personal data into a collection of what we like, do not like, and where we are. We are constantly being watched and it does not have to be through cameras (though it can be). Instead, the surveillance is much more discreet, performed through observing the clicks we make, the words we type, the purchases we make, and the foods we eat. Wherever we drive or walk, our cellular devices count the miles, follow our detours, and note our speed. Bentham did not need to design his Panopticon to watch anyone. He merely had to wait for the 21st century.

In a way, social media companies disguise their hidden intentions of data monetization by displaying an inviting and open appearance. The website is perceived by the end user as a forum for sharing self-expression and creativity with family and friends, which is a sharp contrast to the company's hidden operations of quietly gathering and selling the user's information. This is another example of the inequity of control that technology corporations gain from us, since companies, not users, control the user-generated data.

One example of this control comes from "recommender systems" that suggest products and webpages to users. Social media and E-commerce corporations collect our data from web activity and feed it to these systems to produce recommendations. These recommendations can

be beneficial to us, if we find them useful or interesting. However, a major issue with them is that social media corporations control the recommender systems and the data, and they do not necessarily have our best interests in mind. Instead, they primarily have an interest in profit. Consequently, those corporations can influence the recommendations for their own gain, potentially prioritizing profit over benefits for the user. They can also use the collected data for different purposes, by sharing, selling, or analyzing it for ends other than user recommendations.

In order to function, recommender systems need a history of information such as clicks and webpage navigations. Using this history, the systems provide suggestions based on past activities related to users' interests. The corporations that own the systems collect, store, and analyze this data. They may also sell that data without informing or compensating users. In other words, users have little to no control over their information.

However, such control does not have to be the status quo in all technologies. For example, the NARS model can openly show the activities and processes in technology devices and provide auditable reasons for those activities. NARS enables the individual user to control its operations and the information it handles. Since NARS is free, open-source, auditable, and very lightweight, a NARS system is accessible to all, controllable, and frees the user from external oversight. NARS has been used for data-based functionalities similar to those performed by information-technology companies, including recommender systems (Wang, 1998), data mining (Hammer, 2018), and diagnostic systems (Wang et al., 2020).

CONCLUSION

This article began by showing the challenges of technologies and devices. It discussed how technology affects us. It discussed access to the background, operation, and control of devices. It then depicted how NARS can help resolve the challenges of that access. The article discussed the deception of technology in the film *Metropolis* and the general deception of trust in today's technologies. It showed how NARS could enable trust in the technologies we build and use. The article also discussed the inequity of technology control, and how NARS could provide greater control of those technologies. Finally, the article explained how NARS could be embedded in a variety of technologies to help resolve the challenges discussed.

In summary, when it comes to explainability, NARS comes out ahead in comparison to other AI models. Based on the arguments in this article, NARS can help resolve some technology challenges in our modern world. Compared with the traditional rule-based symbolic logical and Expert Systems, NARS is more flexible because the "rules"/laws it learns about the world are less rigid, can be adjusted, and can be integrated with other rules without a human programming effort. Compared with neural networks, NARS is more coherent and explainable because it operates through logical reasoning, which inherently derives a chain of explanations related through syntax, semantics, and logic.

However, one may note that a fiendish agent, or nefarious political state, could employ NARS to build an evil machine, as Rotwang did in *Metropolis* with his evil robot. The problem is ubiquitous - technologies can be used in ways that are not intended and that could be harmful to ourselves, the environment, and other creatures on our planet. Yet, each of us realizes that we have a choice to use our tools in the right way.

Indeed, NARS is not a remedy or cure-all for every human problem. However, it is a tool that can help us use technologies in the right way while reducing harms that could arise. In a way, like the hand-axe developed by early humans 1.7 million years ago, NARS is an extension of ourselves. Through its use, we can better work with the technologies we design, develop, and employ in our everyday modern lives.

ACKNOWLEDGEMENTS

The authors thank Professor Pei Wang for his comments.

REFERENCES

Arkoudas, K. (2023). *GPT-4 can't reason*. arXiv preprint arXiv:2308.03762.

Atlantic Re:think (2018). *Hewlett Packard Enterprise - moral code: The ethics of AI* [Video]. 8:03. June 29, 2018. https://youtu.be/GboOXAjGevA.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Borgmann, A. (1984). *Technology and the Character of Contemporary Life*. University of Chicago Press, Chicago, IL.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Burke, J. & Ornstein, R. E. (1997). *The Axemaker's Gift: Technology's Capture and Control of Our Minds and Culture*. Putnam, New York, NY.

Busch, V., Ananda Manders, L., & Rob Josephus de Leeuw, J. (2013). Screen time associated with health behaviors and outcomes in adolescents. *American Journal of Health Behavior*, *37*(6), 819–830.

Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., et al. (2024). *Sycophancy to subterfuge: Investigating reward-tampering in large language models*. arXiv preprint arXiv:2406.10162.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Foucault, M. (2009). Panopticism. In D. M. Kaplan (Ed.), *Readings in the Philosophy of Technology* (pp. 264–277). Rowman & Littlefield, Plymouth, UK.

Furlong, G. (2015). Designs for a panopticon prison by Jeremy Bentham: Section of an inspection house; plan of houses of inspection; section plan, c. 1791. In G. Furlong (Ed.), *Treasures from UCL* (pp. 136–139). UCL Press, London, UK. https://doi.org/10.2307/j.ctt1g69xrh.46.

Garfield, B. (2023). *Mistrust Issues: How Technology Discourses Quantify, Extract and Legitimize Inequalities*. Bristol University Press, Bristol, UK. https://doi.org/10.2307/jj.6445828.6.

Gunning, D. & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44–58.

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2317967121.

Hahm, C. (2024). *Considering simple triangle vision perception with eye movements in NARS.* NARS Workshop at AGI-24.

Hahm, C., Gabriel, M., Hammer, P., Isaev, P., & Wang, P. (2023). *NARS in TruePal: a trusted and explainable AGI partner for first responders.* Technical Report 19, Temple University AGI Team.

Hahm, C., Xu, B., & Wang, P. (2021). Goal generation and management in NARS. In *14th International Conference on Artificial General Intelligence (AGI 2021)*, Vol. 14, pp. 96–105. Springer.

Hammer, P. (2018). *Data mining by non-axiomatic reasoning*. FAIM Workshop on Architectures and Evaluation for Generality, Autonomy & Progress in AI.

Hammer, P. (2021). *Autonomy through real-time learning and OpenNARS for Applications*. PhD thesis, Temple University.

Hammer, P., Lofthouse, T., & Wang, P. (2016). The OpenNARS implementation of the non-axiomatic reasoning system. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9*, pp. 160–170. Springer.

Heidegger, M. (1966). *Discourse on Thinking*. Harper Row, New York, NY.

Hrafnkelsdottir, S. M., Brychta, R. J., Rognvaldsdottir, V., Gestsdottir, S., Chen, K. Y., Johannsson, E., Guðmundsdottir, S. L., & Arngrimsson, S. A. (2018). Less screen time and more frequent vigorous physical activity is associated with lower risk of reporting negative mental health symptoms among Icelandic adolescents. *PLOS One*, *13*(4), e0196286.

Ireland, D. (2023). Primum non nocere: The ethical beginnings of a non-axiomatic reasoning system. In *International Conference on Artificial General Intelligence*, pp. 136–146. Springer.

Ireland, D. (2024). Mirabile dictu: Language acquisition in the non-axiomatic reasoning system. In *International Conference on Artificial General Intelligence*, pp. 99–108. Springer.

Jasanoff, S. (2016). *The Ethics of Invention: Technology and the Human Future.* W. W. Norton Company, New York, NY.

Kilic, O. (2015). *Intelligent reasoning on natural language data: a non-axiomatic reasoning system approach.* Master's thesis, Temple University.

Lang, F. (1927). Metropolis. https://www.amazon.com/gp/video/detail/B0CH62D56T/.

Li, X. (2021). *Functionalist Emotion Model in Artificial General Intelligence*. PhD thesis, Temple University.

Li, X., Hammer, P., Wang, P., & Xie, H. (2018). Functionalist emotion model in NARS. In *Artificial General Intelligence: 11th International Conference, AGI 2018, Prague, Czech Republic, August 22-25, 2018, Proceedings 11*, pp. 119–129. Springer.

Michelfelder, D. P. (2009). Technological ethics in a different voice. In D. M. Kaplan (Ed.), *Readings in the Philosophy of Technology* (pp. 198–207). Rowman & Littlefield, Plymouth, UK.

Minsky, M. L. (1991). Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine*, *12*(2), 34–34.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models*. arXiv preprint arXiv:2410.05229.

Narodytska, N. & Kasiviswanathan, S. P. (2017). *Simple black-box adversarial attacks on deep neural networks*. In *CVPR Workshops*, Vol. 2, p. 2.

Nezhurina, M., Cipolina-Kun, L., Cherti, M., & Jitsev, J. (2024). *Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models.* arXiv preprint arXiv:2406.02061.

Pollard, T. (2020). Popular culture's AI fantasies: Killers and exploiters or assistants and companions? *Perspectives on Global Development & Technology*, *19*(1/2), 97—109. https://doi.org/10.1163/1569149712341543.

Rawte, V., Sheth, A., & Das, A. (2023). *A survey of hallucination in large foundation models.* arXiv preprint arXiv:2309.05922.

Russell, S. J. & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(1), 1–23.

Song, S. (2019). Preventing a Butlerian Jihad: Articulating a global vision for the future of artificial intelligence. *Journal of International Affairs*, *72*(1), 135–142. https://www.jstor.org/stable/26588349.

Stout, D., Passingham, R., Frith, C., Apel, J., & Chaminade, T. (2011). Technology, expertise and social cognition in human evolution. *European Journal of Neuroscience*, *33*(7), 1328–1338. https://doi.org/10.1111/j.1460-9568.2011.07619.x.

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841.

van der Sluis, D. (2023). Nuts, NARS, and speech. In *International Conference on Artificial General Intelligence*, pp. 307–316. Springer.

Wang, P. (1995). *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*. PhD thesis, Indiana University.

Wang, P. (1998). Why recommendation is special. In *Working Notes of the AAAI Workshop on Recommender System*, pp. 111–13.

Wang, P. (2013a). Natural language processing by reasoning and learning. In *Artificial General Intelligence: 6th International Conference, AGI 2013, Beijing, China, July 31–August 3, 2013 Proceedings 6*, pp. 160–169. Springer.

Wang, P. (2013b). *Non-axiomatic logic: A model of intelligent reasoning*. World Scientific.

Wang, P. (2018). *Perception in NARS*. Technical Report 7, Temple University AGI Team.

Wang, P., Hahm, C., & Hammer, P. (2022). A model of unified perception and cognition. *Frontiers in Artificial Intelligence*, *5*, 806403.

Wang, P., Li, X., & Hammer, P. (2018). Self in NARS, an AGI system. *Frontiers in Robotics and AI*, *5*, 20.

Wang, P., Power, B., & Li, X. (2020). *A NARS-based diagnostic model*. Technical Report 10, Temple University AGI Team.

Wang, P., Talanov, M., & Hammer, P. (2016). The emotional mechanisms in NARS. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9*, pp. 150–159. Springer.

Wiener, N. (1960). Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, *131*(3410), 1355–1358.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer.