Ethical Intelligence: Evaluating Transparency, Fairness, and Accountability in AGI through the NARS Framework

Artificial General Intelligence Project
Nitin Vadnala, Pranav Nemani
May 2, 2025

Abstract

As Artificial General Intelligence (AGI) is quickly advancing, the lack of seeing the need for ethical frameworks to guide development is worrying. This paper takes a look at the ethical implications with a focus on transparency, accountability, and fairness. Non-Axiomatic Reasoning System (NARS) provides a transparent framework with a level of reasoning that runs into challenges at an AGI level complexity. This calls for the need of Explainable AI (XAI) and visualization tools. Fairness issues in this model are apparent due to the input knowledge learning and prioritization which calls for bias detection and fairness checks on the Non-Axiomatic Logic (NAL). Accountability requires adapting current frameworks intertwined with human oversight and feedback loops. The paper pinpoints collaboration as necessary for ethically responsible AGI development using the NARS framework.

I. Evaluating Ethical Implications in AGI

Artificial intelligence and its rapid growth has sprouted the concept of Artificial General Intelligence (AGI) and has made it a tangible goal to attain for researchers and engineers. With AI systems showing convincing capabilities equal to those of a human or even further, the ethical implications of such profound intelligence systems becomes an important topic at hand. The possibility of AGI revolutionizing how machines learn, think, and act necessitates the parallel development of ethical guidelines to guide its creation and deployment. Fundamental ethical challenges in AGI development span dozens of important areas, from ensuring safety and control of autonomous systems to addressing ubiquitous challenges of bias and fairness, managing the possibilities of extensive economic and workforce transformation, safeguarding privacy and preventing unwarranted surveillance, mitigating existential risks of advanced AI, and developing requisite transparency and accountability mechanisms. Interdisciplinary collaboration is needed to effectively address these intricate challenges and ensure that AGI development aligns with human values and priorities.

One of the prevalent paths that lead to the strong development of AGI is the Non-Axiomatic Reasoning System. NARS is an adaptive reasoning system that can learn and operate under a

real AGI-world constraint of limited knowledge and resources. This system simulates cognitive powers and makes use of experience based semantics. This is so that meaning and truth of the concept are a function of how it has developed in the world, and that evidence is available to the system. The architecture of NARS, particularly its AIKR basis and its use of experience for reasoning basis, suggests that its ethical implications should be perceived rather differently from those of other AI paradigms. This report evaluates the ethical implications of AGI, particularly within the NARS architecture, bringing to attention the important dimensions of transparency, accountability, and fairness. By discussing the conjunction of NARS's architecture and principles with ethical matters, this analysis aims to create further understanding of the opportunity to develop ethically responsible AGI.

II. Understanding the Non-Axiomatic Reasoning System

NARS operates under an umbrella of core principles that make it distinct from conventional systems and many other AI infrastructures. At its foundation is the principle that intelligence is the capacity of a system to survive in the environment and work effectively even in the absence of sufficient knowledge and resources, a principle formalized as the Assumption of Insufficient Knowledge and Resources (AIKR). AIKR states that NARS possesses limited processing capacity, must operate in real time, remains open to the potential for unexpected tasks, and acquires expertise from experience. This is the perspective on intelligence advocated by our Professor Pei Wang, which recalls a sense of relative rationality where the validity of conclusions will be determined with respect to available knowledge and resources within the system rather than relative to an absolute truth.

Being built with the inclusion of AIKR is the idea of experience-based semantics. In NARS, the truth-value of a statement is quantified as the amount of evidence for it in regards to the system's experience, where the meaning of a symbol is characterized as connections to other symbols. This is different from how model-theoretic semantics operate, where sufficient knowledge is utilized to convert symbols into an objective form. Further, NARS is rational, for example, its conclusions are the closest it can provide as per its existing constraints. The system is non-axiomatic in the sense that all beliefs are revisable upon the attainment of new evidence. These processes are completed through a single mechanism of reasoning based learning. The premise of AIKR impacts NARS' knowledge acquisition and processing to produce a more responsive system, while simultaneously posing issues regarding the reliability and bias created through biased or limited experiences.

NARS architecture consists of a dynamically structured memory as a network. It contains a single, uniform reasoning-learning process while running various cognitive processes. It executes asynchronously and in parallel. It uses a task pool and knowledge base with tasks and knowledge to be acquired for processing based on the priority distribution. To deal with

limitations of AIKR, NARS has a characteristic, a mechanism of forgetting allowing it to handle a limited amount of storage. NARS' emphasis on parallelism and dynamic memory enables it to deal with complexity, however introduces problems of predictability and ethical manipulation of its emergent behavior as well.

The non-axiomatic logic aspect of NARS is the Non-Axiomatic Logic (NAL). NAL is a term used to express and handle various forms of uncertainty. NAL phrases typically take the form "Subject + Predicate <f;c>", where 'r' is an inheritance relation and '<f;c>' is a truth value that consists of frequency and confidence. NAL provides a unified formalism for many types of inferences, including deduction, induction, abduction, and revision. NAL makes use of local, forward, and backward inference rules, whose strength is dependent on the confidence degree of the conclusions. NAL semantics are experience-based and try to determine the truth value of the statement which is dependent on evidential support from the system's experience. This distinction from traditional reasoning, with AIKR, sparks questions about the danger of biases being a part of and spreading throughout the system's inferences due to the nature of its previous experience.

III. Ethics in Artificial General Intelligence

Transparency is a very important aspect of AI ethics as it acts as a building block during the stage of creation and usage. With transparency comes a strong commitment to trust and responsibility. In the eye of AGI, transparency points to being aware with the stakeholders of showcasing how and why the system makes a decision and output. To guarantee accountability, losing a level of transparency can remove how the public trusts the system and create a massive hurdle to overcome. Frameworks that aid in regulation such as the European Union's GDPR and AI Act require openness from the AGI system which shows how important transparency is legally and within society. With all that being said, it's easier said than done to be transparent within a complex model like AGI. These systems have a sense of holding information outside of the box where these machine learning models can’t exactly be interpreted simply.

Having accountability in AGI makes sure that developers and users are held responsible for what is outputted. Without a clear structure on accountability, it's difficult to point out failures. The EU AI Act sets clear guidelines that allows the development of the system to be held responsible. Establishing accountability for the level of autonomy of AGI brings up challenges where deciding on what the system is responsible for is a very complex operation. Creating an oversight mechanism is essential for the development of AGI to stay transparent in research, risks, and development. All of which, in the end, creates a foundation of accountability.

Fairness in AGI systems points out concerns about bias and discrimination. These are very much apparent issues seen in many other AI applications. Bias in AI can source from many aspects of biased training data and design which in return leads to discrimination outcomes.With the guarantee of fairness in AGI, individuals are made sure of being treated equitably as the system takes human rights and justice into consideration. Algorithms face limitations in equitable outcomes in a diverse population without regulation and an attempt to limit bias. Addressing harm sourcing from unfair AI systems requires an intricate understanding of the sources of bias and base implementation to be able to accurately mitigate throughout the AGI lifecycle.

IV. Transparency

NARS has built-in intrinsic features which aid in its ability to be transparent, especially when compared to AI systems in modern society which incorporate “black box” setups. The knowledge representation in symbolic form, with the use of NARS language, enables the readability to be human-like for beliefs and goals. In addition the reasoning process that NARS utilizes works through the use of inference rules defined in NAL, which provides a retractable pathway from the system's initial knowledge all the way to the derived conclusions. The truth value of every statement in NARS is a second source of transparency through its reporting the amount of evidential support for the statement, quantifying the system's belief in what it knows. These characteristics mean that the underlying structure of NARS allows more insight into its internal workings and decision-making than models whose reasoning and knowledge are encoded in complex, sub-symbolic structures like the weights of a neural network.

Although consisting of inherent transparency circumstances, AGI and its potential to achieve complete transparency comes with its own set of challenges. With the growth of NARS’ knowledge base and its inference chains becoming complex, the difficulty to fully understand the systems reasoning as humans. The dynamic nature of NARS's memory and task prioritization mechanisms can also interfere with reconstructing the precise chain of reasoning that led to a particular conclusion. Moreover, NARS's experience-grounded semantics lead to the "meaning" of words being internal to the system, being a function of its own past interactions, rather than always exactly as intended or comprehended by humans. The individual components of NARS, such as its rules and representations of knowledge, can be transparent in themselves. The emergent behavior arising from their complex interactions within a system scaled to AGI levels might still present a form of opacity to human observers.

To continue the development of NARS, explainable artificial intelligence methods can be utilized and applied to reasoning systems that operate symbolically. There are fields of AI which combine the abilities from pattern recognition in neural networks as well as the logic aspect from symbolic systems, which can aid in the pathway to interpretable AI. As an example

PyReason which is an explainable AI system, promises the potential for supporting inferences with explainability in symbolic systems. Such techniques applied to NARS may produce more human-readable explanations of its reasoning processes, effectively translating the system's internal logic into more human-comprehensible terms. This could involve summarizing the most significant premises and inference steps that played the greatest role in deriving a particular conclusion.

With the use of visualization tools, we can better understand NARS' internal states and processes. Due to NARS' memory being interpreted in the form of a network, its knowledge base as a graph of concepts and relationships being visualized can provide clarity into the system's perception. Likewise, flowcharts of rule applications or tree representations of the systems reasoning could aid in tracing steps that NARS takes to derive a conclusion. Such as these, tools like semantic networks and visual knowledge graphs provide an opportunity to observe and apply NARS' memory structure, proving to be a great track to proceed on to further improve transparency and behavior auditing.

V. Fairness and Bias

Many aspects of the Non-Axiomatic Reasoning System (NARS) showcase areas of bias. Specifically, this is seen in the areas where it reasons and makes inferences. The foundation of NARS is built strictly from the data it is given. In simple terms, this means that the environment that the system bases its output from is coming from what we only provide the system, nothing else. If the input data ends up being biased or not being accurate, the system's rationale and beliefs will now show that bias. This can lead to a skewed result where the inferences for concepts and relationships are not aligned with the truth. The AIKR principle allowed for a more limited and selective memory process. It is capable of now incorporating a priority queue, in a sense, where it can put some beliefs and inputs over others. This selective decision making can also provide bias to the system if the priority is given to certain types of information consistently. The probability of the selected concepts or tasks is based on a priority value. However, this leads to the bias of the initial priorities swaying towards particular pieces of information. Even the truth value functions of NARS's inferences can favor certain areas of information which lead to a biased output as well. The system's inherent operation of any starting point knowledge can also bring up inherent bias. This shapes the way the system learns and reason. The emotional aspect of NARS is another source of bias where a positive or negative connotation is given to certain concepts. This, once again, can cause a sway towards one side or another in the output. The process of NARS inferring relationships from its experience is easily biased in causal reasoning.

NARS priority method can impact how bias plays out. The priority method for processing tasks could cause certain areas of the system's knowledge to be more descriptive and more

explored. This occurs when NARS is dynamically adapting the priorities for given information. The initial priority level given to a task could favor a said conclusion that supports a belief. This is based on the frequently used reasoning pattern that the system has already adapted to an extent. However, tasks that are related to older or less used knowledge can lose priority if we implement a forgetting aspect where bias is created against past used relevant information. NARS tends to struggle to allocate its resources to more of a blanket of information for a balanced understanding. As it works based on the task priority, it tends to overlook certain tasks that are important for a balance. The timing of user inputs can impact the prioritization process which will inadvertently shift the system's focus on certain areas.

Possible biases can also come up from term logic. This is created from the basis of NARS;s knowledge foundation. Since the meaning of terms in NARS is pulled from its knowledge and inputs, a limited number of inputs can create a biased understanding. The creation of a relationship implementation between terms based on frequency that they show up in the system's knowledge can highlight biases in the dataset. The truth values linked to each input in NARS is based on the number of positive and negative evidence that is seen. This puts them at risk of bias once again in how it is collected and processed. Even though beliefs can be edited, initial bias seen by the system can be hard to overcome. This then requires the system to introduce counter arguments to balance out the model's knowledge.

In order to aid the system to reduce the risks of bias in NARS, a multi step approach will be needed to fully mitigate this. Starting off the system's experiences and training has to cover the entire basis. Meaning, it can't hone in on a specific sector of its knowledge and needs to stay diverse and cover all areas of its knowledge that the system is being built on. Possibly by applying already existing bias detection models or creating new ones will be very beneficial for pointing out and eliminating biases within the NARS's knowledge and reasoning. Researching fairness limitations to integrate directly into the Non-Axiomatic Logic (NAL) could aid in carving a path for NARS to apply a well rounded level of reasoning. As an extra level of protection, incorporating a human-like mechanism to be able to detect and evaluate the bias corrections from the output could be a beneficial extra wall. Continuously monitoring the behavior of the system for signs of bias is essential for the system to be able to consistently work fairly during all aspects of the life cycle.

VI. Accountability

Being able to gauge accountability for how the AGI system acts within NARS requires a bit of attention. Current frameworks developed for AI accountability creates a foundation to hold NARS to at the very least. These frameworks highlight the need for information flow. This refers to documents and tracking the AI system's development and integration journey. On top of this, the relevant stakeholders within the project are also accountable. They focus on how

essential it is to perform system evaluations periodically which include assessments and certifications on these high level, high risk applications. These frameworks downplay the role of the government and important bodies in developing an accountability ecosystem. With this ecosystem comes a level of standard to be met, practices checks, and actually resource allocation into the research to further the ecosystem. Risk and impact assessments are massive building blocks to demanding an accountability framework. This requires developers to be proactive and manage the possible risks linked with AI. Key components include testing, transparency, high quality data, and strong security measures where all these parts play a role in guaranteeing AI systems are kept within a controllable reach. Bringing in individuals such as AI Risk Officers are also often considered to guarantee responsible AI development and creation.

Applying these accountability frameworks to an AGI system like NARS has a few hurdles to overcome simply due to the high level of autonomous interactions the system undergoes. It's essential to build mechanisms that guarantee that NARS can work in line with human values and ethics. Given the high ceiling of advancing AGIs and how they engage with self replication, transparency in these processes are important to meet. As these AGI systems become more and more complex, questions about their moral levels and rights need to be considered in the idea of accountability. Applying traditional measures to these systems with advanced reasoning and capabilities will likely need more of a keen eye. These systems will require a more out of the box approach to go beyond the current practices which, for example, can consider AGI's understanding and its level of response to ethical viewpoints.

Human oversight will play a key role in making sure that NARS meets a level of ethical operation status. Human intervention is important for ethical decision making as it provides a moral gauge that AI systems tend to lack. Humans can define a set of ethical guidelines and boundaries for NARS alongside a review process to validate if an output meets a standard or not. This is done, more importantly, in more critical, complex applications where the possibility of crossing a boundary is more possible. The creation of accountability guidelines for the actions of NARS is put directly into the hands of individuals who are developing, deploying, and overseeing the usage. Human oversight also aids the AI system where it struggles in navigating dynamic situations where more adaptability or contextual understanding is needed. By incorporating feedback and timely adjustments, humans can watch the continuous cycle of improvements on a NAR system's ethical actions. Implementing intervention points allows humans to review and override NAR's actions in certain situations where it can act as a safety net.

Incorporating user feedback cycles into the development of NARS is a key puzzle piece to include to shape the AGI system responsibly. Ethical feedback loops can allow users to highlight concerns, biases, and suggest any improvements to elevate their experience. This

would allow for the system to be more responsive and adaptive to the environment that the users' value and need. The benefit of deploying this would allow for improved accountability, increased transparency, stronger trust in the system, and continuous improvement. Key components to include within the design language to be effective within the ethical feedback chain would include accessibility, transparency about how feedback will be used, guarantee privacy for users, and providing timely responses to feedback. Implementing a feedback channel where AI can be used to process the feedback given can aid in quickening the developmental changes. Continuous monitoring and improving upon these systems are important strategies for the creation of an ethical feedback loop for AGI.

VII. Discussion

The progression of AGI through the NARS structure shows a very complex path forward between its basic design principles and the ethical challenges of transparency, accountability, and fairness. NAR's depends strongly on Assumption of Insufficient Knowledge and Resources (AIKR) and experience based conclusions which sheds some light in the tunnel for a very adaptable level of intelligence. However, with these comes a set of ethical hurdles. While AIKR allows NARS to be able to function in a real environment with minimal information, it also correlates to the system working in a biased setting with incomplete knowledge of the entire picture. This makes the model's understanding and decision making tip on to the bias scale. Experience based approaches allow NARS to evolve its understanding which can lead to experience impacted knowledge which may not align with human values.

The transparency that NAR's showcases within its reasoning rules provides a massive level up over many other symbolic AI systems. The ability to dive deep into the knowledge basis and trace an inference allows a level of understanding that can be provided. This can aid in overseeing where ethical issues may arise and how to mitigate it. The complexity of NARS moving towards AI still causes many challenges to come about. With such a massive amount of knowledge networks and reasoning pathways, this can still hinder the model's decision making process. This challenge can be overcome with the use of Explainable AI (XAI) and visualization techniques.

Creating a baseline of accountability for NARS will ask for an adaptation of existing AI accountability frameworks for an AGI system to be able to run under AIKR. This requires the development of approaches to risk and impact assessment, testing, and transparency, all of which must be altered based on NAR's learning and reasoning structures. Human oversight is still a major role player where they can provide ethical guidance and review input and output to guarantee that the model is in line with human values. Incorporating user feedback through ethical feedback cycles can provide a real-world angle to impact NARS's ethical performance. This will not only help point out unseen issues but also aid in an approach to mitigate as well.

NARS having a very unique architecture and reasoning approaches showcases potential for developing ethical frameworks for AGI. It's capable of learning and reasoning based off of experience alongside a level of transparency, it has the perfect foundation for experimenting with different approaches. These include applying ethical principles and gauging their effectiveness to an uncertain environment. The implications of applying an AIKR approach to AGI for the future of AI ethics research is a great step forward. It provides a shift towards designing a unique intelligent model that realizes its own shortcomings and looks to adapt and learn efficiently to stay within ethical limits.

VIII. Conclusion

This report took a look at the ethical implications of Artificial General Intelligence (AGI) through the model of the Non-Axiomatic Reasoning System (NARS) which focused on the essential aspects of transparency, accountability, and fairness. This analysis showcases that even though NARS does have a level of transparency with its simple nature of reasoning, reaching full transparency in a scaled AGI will require the use of Explainable AI (XAI) methods and visualization tools. Guaranteeing fairness in the model needs a multi angular approach to mitigate possible bias that surfaces from its experienced based learning and long alongside its task prioritization. Integrating a fairness constraint into the core logic can allow for this to go a step further. Setting a baseline for accountability for NARS requires an adaptation of an existing AI accountability framework to the challenges seen in Assumption of Insufficient Knowledge and Resources (AIKR). The final building block to this model would be ensuring the major role that human oversight and user feedback is intertwined with the final integration.

To further the ethical development of AGI with the NARS framework, future research and implementation efforts need to hone in on reasoning processes through XAI techniques and visualization methods. Creating a continuous cycle of searching bias within NARS's learning and reasoning habit is important. Pairing this with the development and implementation of bias mitigation approaches tailored to the complex, unique architecture that the model holds. Exploring integration methods of ethical principles into Non-Axiomatic Logic (NAL) could provide a guide to NARS towards ethical behavior.

Navigating the ethical challenges of AGI to be responsible in the development process of systems like NARS asks for a strong level of collaboration. AI researchers must work with ethicists, policymakers, and even the public to guarantee a shared understanding of the ethical issues that arise and to build effective guidelines. The NARS framework, as a unique approach to AGI, offers a strong foundation for the understanding of these said challenges and exploring possible solutions that can aid the development of ethically focused AGI for the benefit of the public.

References

Adah, William Arome, et al. “The Ethical Implications of Advanced Artificial General Intelligence: Ensuring Responsible AI Development and Deployment.” SSRN Electronic Journal, 2023, https://doi.org/10.2139/ssrn.4457301.

Amodei, Dario, et al. “Concrete Problems in AI Safety.” arXiv Preprint, 2016, https://arxiv.org/abs/1606.06565.

Brundage, Miles, et al. “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.” arXiv Preprint, 2018, https://arxiv.org/abs/1802.07228.

Doshi-Velez, Finale, and Been Kim. “Towards a Rigorous Science of Interpretable Machine Learning.” arXiv Preprint, 2017, https://arxiv.org/abs/1702.08608.

Floridi, Luciano, and Josh Cowls. “A Unified Framework of Five Principles for AI in Society.” Harvard Data Science Review, vol. 1, no. 1, 2019, https://doi.org/10.1162/99608f92.8cd550d1.

Ireland, David. “Primum Non Nocere: The Ethical Beginnings of a Non-Axiomatic Reasoning System.” Lecture Notes in Computer Science, 2023, pp. 136–146, https://doi.org/10.1007/978-3-031-33469-6_14.

Jobin, Anna, Marcello Ienca, and Effy Vayena. “The Global Landscape of AI Ethics Guidelines.” Nature Machine Intelligence, vol. 1, no. 9, 2019, pp. 389–399, https://doi.org/10.1038/s42256-019-0088-2.

Nabil, Ashrafur Rahman, et al. “Ethical Implications of AI-Powered Predictive Policing:

Balancing Public Safety with Privacy Concerns." Innovatech Engineering Journal, vol. 2, no. 1, 2025, pp. 47–58, https://doi.org/10.70937/itej.v2i01.54.

Sedat Sonko, et al. "A Critical Review Towards Artificial General Intelligence: Challenges, Ethical Considerations, and the Path Forward." World Journal of Advanced Research and Reviews, vol. 21, no. 3, 2024, pp. 1262–1268, https://doi.org/10.30574/wjarr.2024.21.3.0817.

Wang, Pei. "On Defining Artificial Intelligence." Journal of Artificial General Intelligence, vol. 10, no. 2, 2019, pp. 1–37, https://doi.org/10.2478/jagi-2019-0002.