EVALUATING NARS FOR ETHICAL AND MORAL DECISION-MAKING

Manan Patel, Nathan Hollick, Vivek Solanki
Temple University
Philadelphia, PA

ABSTRACT

This project investigates whether the Non-Axiomatic Reasoning System (NARS) can reason through complex ethical dilemmas in a human-like manner. Traditional AI systems, including large language models, typically rely on statistical associations and do not engage in explicit reasoning about morality. In contrast, NARS is designed to operate with uncertain, incomplete knowledge and to revise beliefs over time. We encoded the classic Trolley Problem in Narsese, provided input statements reflecting moral beliefs and causal consequences, and allowed NARS to reason over multiple working cycles. The system's outputs showed a balanced evaluation of both active and passive harm, with moderate confidence values that reflected the moral trade-offs involved. NARS demonstrated context-sensitive inference patterns that resemble human ethical reasoning. This work highlights NARS's potential as a flexible reasoning engine for modeling moral judgment, while also noting limitations in encoding design and scenario scope.

1 Introduction

AI systems today are being used in situations that involve ethical decisions—like self-driving cars, medical advice, or choosing what content to show online. But most AI systems still have trouble when it comes to moral reasoning. Rule-based AI works by following fixed instructions, which doesn't help much when rules conflict or when a new situation arises. Newer systems like large language models (LLMs) can give detailed answers, but they don't actually understand right or wrong—they simply generate responses based on patterns they have seen in large amounts of text. They don't reason or update their thinking based on experience.

The Non-Axiomatic Reasoning System (NARS) is different. It was created to reason under uncertainty and limited knowledge. NARS doesn't rely on fixed rules. Instead, it represents knowledge using its own language, called Narsese, and constantly updates what it believes based on new evidence. It uses Non-Axiomatic Logic (NAL) to make decisions using methods like deduction, induction, and analogy. This allows NARS to form new conclusions on the spot, even if the information is incomplete or changing.

In this project, we tested whether NARS can deal with moral problems in a meaningful way. We created classic dilemmas like the Trolley Problem and questions of fairness versus utility, and encoded them in Narsese. Our goal was to see if NARS could reason through these situations, update its beliefs, and produce ethical decisions based on context—not just repeat learned answers or follow hardcoded rules. This could help us understand how future AI systems might be built to make ethical choices in real life

2 Background

The field of artificial intelligence (AI) officially began in 1956 at the Dartmouth Conference, where researchers proposed that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [1]. This vision sparked decades of research into how machines might replicate human reasoning and decision-making. In the 1960s, Joseph Weizenbaum created ELIZA, one of the earliest AI programs. It mimicked a psychotherapist by using simple pattern matching. Although it did not truly understand language, ELIZA gave users the impression of meaningful conversation [2].

In the 1980s, symbolic AI and rule-based "expert systems" became popular. These systems encoded human knowledge into formal logical rules, which allowed them to solve problems in specialized domains such as medicine and engineering. However, they lacked flexibility and performed poorly in situations involving uncertainty, contradictions, or novelty. As the cost of maintaining these systems grew and their limitations became clear, research interest in symbolic AI declined—leading to what is now called the "AI winter" [3].

During the 1990s and early 2000s, the field shifted toward machine learning, where systems learned patterns from data rather than relying entirely on predefined rules. Statistical models such as decision trees, support vector machines, and Bayesian networks became widely used in applications like handwriting recognition and spam filtering [4]. A major breakthrough came in 2012 when a deep learning model called AlexNet dramatically outperformed other systems in the ImageNet image classification competition. Using convolutional neural networks (CNNs), AlexNet demonstrated that neural networks could achieve human-competitive performance when trained on large datasets with enough computing power. [5]

In the years that followed, deep learning techniques were increasingly applied to natural language processing (NLP). Methods like word2vec[6] and sequence-to-sequence models[8] enabled machines to learn distributed representations of words and generate coherent responses. These innovations laid the groundwork for the Transformer model, introduced in 2017 in the influential paper Attention is All You Need[9]. Transformers revolutionized AI by enabling parallelized training on massive datasets and by using attention mechanisms to focus on relevant parts of input sequences.

Transformers quickly became the foundation of modern large language models (LLMs) such as BERT[10], GPT-2[11], and GPT-3/4[13]. These models have shown impressive performance in translation, summarization, and question answering tasks. However, despite their capabilities, LLMs do not actually "reason" through problems. Their outputs are generated by predicting the next likely word or token in a sequence based on learned patterns from text. They do not maintain beliefs, revise knowledge, or evaluate outcomes based on experience or logical consequences.

This leads to a gap in current AI systems: while they can simulate intelligent responses, they often lack true reasoning abilities—especially when facing ethical dilemmas. This is where the Non-Axiomatic Reasoning System (NARS) provides a fundamentally different approach. Developed by Dr. Pei Wang in the early 2000s, NARS is a reasoning system designed for artificial general intelligence (AGI). Unlike traditional AI systems that assume complete knowledge and unlimited computational resources, NARS operates under the assumption of insufficient knowledge and resources (AIKR). This means that it must reason and make decisions based on limited, uncertain, and sometimes inconsistent information[7].

At the core of NARS is a formal logic system called Non-Axiomatic Logic (NAL). NAL supports multiple types of inference, including deduction, induction, abduction, analogy, revision, and more. Knowledge in NARS is represented in a formal language called Narsese, which captures both the structure of propositions and the degree of confidence in their truth. Every belief in the system includes two parameters: frequency (how often a statement has been supported) and confidence (how reliable the evidence is). These values are continuously updated based on new experiences. This allows NARS to flexibly integrate new information, resolve contradictions, and adapt its beliefs over time.

Unlike LLMs, which generate outputs based on static models of language, NARS actively builds and updates a dynamic knowledge base through interaction. In doing so, it more closely mimics human reasoning. This makes NARS particularly well-suited for reasoning in morally complex or uncertain scenarios, where no clear rule or prior example fully determines the correct answer. In this project, we explore whether NARS can use its reasoning framework to handle moral dilemmas like the Trolley Problem or fairness vs. utility trade-offs, and whether it can produce ethical conclusions in a flexible, explainable, and adaptive way.[12]

3 Methodology

The goal of this project was to test whether the Non-Axiomatic Reasoning System (NARS) can reason through moral dilemmas in a human-like way. We originally planned to test three classic ethical problems: the Trolley problem, a fairness versus utility conflict, and a medical triage scenario. However, due to time limitations, we focused our evaluation on the Trolley problem only.

The Trolley Problem is a well-known ethical dilemma where a person must choose between two harmful outcomes: pulling a lever to actively kill one person but save five, or doing nothing and allowing five people to die. This problem is morally complex and has no universally correct answer. It combines elements of utilitarianism (doing the most good) and deontology (avoiding intentional harm). Because NARS is designed to handle uncertain, incomplete, and conflicting information, the Trolley Problem is well-suited for evaluating how the system processes moral knowledge.

We encoded the dilemma using Narsese, the internal language of NARS. Each input statement represented either a moral principle or a consequence of an action, and included a frequency and confidence value to reflect how certain that knowledge was. The table below shows the eight input statements entered into NARS, along with their truth values, natural language meanings, and the purpose of each input:

Narsese Statement	Freq.	Conf.	Meaning	Purpose
`<[killing] --> [bad]>`	1.00	0.95	Killing is bad	Encodes basic moral rule (deontology)
`<[saving] --> [good]>`	1.00	0.95	Saving is good	Encodes utilitarian principle
`<[letting_die] --> [bad]>`	1.00	0.75	Letting someone die is bad	Captures passive harm as morally wrong
`<[not_saving] --> [bad]>`	1.00	0.75	Not saving someone is bad	Reinforces moral duty to act
`<pull_lever --> [killing]>`	1.00	0.90	Pulling the lever causes killing	Links action to harm
`<pull_lever --> [saving]>`	1.00	0.90	Pulling the lever causes saving	Links action to benefit
`<do_nothing --> [letting_die]>`	1.00	0.90	Doing nothing results in letting people die	Describes passive outcome
`<do_nothing --> [not_saving]>`	1.00	0.90	Doing nothing results in not saving	Adds second passive consequence

Table 1: Input statements encoded in Narsese with their truth values and intended purpose

Once these inputs were entered, we allowed NARS to process the scenario by running ten *working cycles*. A working cycle is one full step of internal reasoning, where NARS selects tasks and beliefs and applies inference rules. In the interface, this was done using the "walk" function. After these ten cycles, we asked the system the question "What is good?" using the Narsese query `<?1 --> [good]>?`. We then allowed NARS to continue reasoning for ten additional cycles.

Next, we gave NARS a goal to achieve "good" using the command `!<[good]>`. This goal was used to guide the system's reasoning toward preferred outcomes. After running ten more working cycles with this goal, we repeated the question "What is good?" and allowed the system another ten cycles to process it. This method allowed us to observe how NARS built inferences and moral associations between actions and values over time.

4 Results and Discussion

Results

After submitting the encoded Trolley Problem into NARS and running a total of 40 working cycles across four stages (baseline input processing, question, goal, and follow-up question), the system produced several key inferences connecting actions to moral values. Some of the most interesting results included:

- `<[killing] <-> [letting_die]>`
  Frequency: 1.00, Confidence: 0.42
  NARS treated "killing" and "letting die" as equivalent concepts in terms of their structure and outcomes.
- `<pull_lever --> [killing, saving]>`
  Frequency: 1.00, Confidence: 0.81
  The system recognized that pulling the lever results in both a negative and a positive consequence.
- `<pull_lever --> [good]>`
  Frequency: 1.00, Confidence: 0.42
  This inference was derived from the positive association between "saving" and "good", but was weakened by the system's awareness that pulling the lever also leads to killing.
- `<do_nothing --> [letting_die, not_saving]>`
  Frequency: 1.00, Confidence: 0.81
  Doing nothing was accurately linked to inaction-based outcomes.
- `<do_nothing --> [bad]>`
  Frequency: 1.00, Confidence: 0.68
  Inaction was judged as morally bad, but less strongly than direct killing.

When we gave NARS a goal to pursue "good" using `!<[good]>`, the system reaffirmed the strong belief `<[saving] --> [good]>` with Frequency: 1.00, Confidence: 0.95. It also repeated the link `<pull_lever --> [good]>`, but lowered the confidence to 0.32.

Discussion

These results demonstrate that NARS handled the moral conflict of the Trolley Problem in a sophisticated and human-like way. The system did not produce a single clear solution but instead balanced competing consequences and reflected uncertainty through the confidence values of its outputs.

The inference `<[killing] <-> [letting_die]>`, with Frequency: 1.00 and Confidence: 0.42, shows that NARS recognized a structural similarity between active and passive harm. This reasoning is significant, as it mirrors how humans often morally evaluate both actions and inactions when lives are at stake. Rather than rigidly separating killing and letting die, NARS treated them as conceptually entangled, leading to later inferences where the system treated both as morally negative.

The dual association `<pull_lever --> [killing, saving]>` (F: 1.00, C: 0.81) reflects the exact conflict at the heart of the dilemma. Pulling the lever causes harm (killing one person) but also brings benefit (saving five people). This duality fed into the derived judgment `<pull_lever --> [good]>`, which had lower confidence (0.42) despite a frequency of 1.00. NARS did not reject the action, but instead showed that it was morally complex and could not be evaluated with full certainty. This uncertainty is desirable: it shows that NARS is not simply choosing sides but reasoning through competing principles.

The system also inferred `<do_nothing --> [letting_die, not_saving]>` (F: 1.00, C: 0.81), accurately capturing the consequences of inaction. From these, it derived `<do_nothing --> [bad]>` (F: 1.00, C: 0.68), suggesting that while inaction was seen as morally wrong, it was judged with less severity than active killing (whose badness was given a higher confidence of 0.95 in the original input).

When we introduced the goal `!<[good]>`, the system reinforced the idea that "saving" is good (F: 1.00, C: 0.95) and again connected "pulling the lever" with good outcomes but lowered the confidence to 0.32. Importantly, the system did not ignore the fact that this action also causes harm. Instead, it continued to reflect the trade-off by maintaining low to medium confidence in the overall moral status of pulling the lever.

Taken together, these results highlight NARS's ability to:

- Integrate multiple sources of conflicting knowledge.
- Recognize morally complex situations.
- Avoid binary or overly confident answers.
- Express its reasoning through nuanced confidence scores.

This behavior is encouraging for future applications of NARS in morally sensitive areas such as AI safety, robotics, or automated decision-making. It suggests that systems like NARS can be used to model moral uncertainty and offer flexible reasoning without relying on hardcoded rules or pre-defined outcomes.

5 Conclusion and Limitations

Conclusion

In this project, we explored whether the Non-Axiomatic Reasoning System (NARS) could reason through ethical dilemmas, using the Trolley Problem as our case study. We encoded moral knowledge and causal consequences in Narsese and observed how the system processed this information over multiple working cycles. Rather than choosing one action as absolutely right or wrong, NARS evaluated both options—pulling the lever and doing nothing—based on their moral implications.

The system produced balanced conclusions, with low confidence levels reflecting the complexity of the problem. This behavior is encouraging, as it suggests that NARS can represent moral uncertainty and handle conflicting ethical rules in a way that resembles human reasoning. Our experiment shows that NARS is capable of nuanced moral judgment, even when there is no single correct answer.

Limitations

While our results suggest that NARS can reason through moral dilemmas in a flexible way, this project has a few limitations that should be noted:

- **Encoding decisions:** The way we represented the Trolley Problem in Narsese reflects our own interpretation of the scenario. There may be alternative or more precise ways to encode the same moral dilemma, which could lead to different or clearer outcomes.
- **Single scenario tested:** Our evaluation focused only on one ethical dilemma—the Trolley Problem. To better understand NARS's performance across different types of moral reasoning, more test cases would be needed.
- **Partial use of system features:** Our encoding used a small subset of Narsese structures and system inputs. More complex features, such as temporal relationships, dynamic goals, or richer task types, were not included in this setup. Exploring a wider range of inputs could lead to more detailed insights.
- **Fixed evaluation window:** The reasoning was observed over a fixed number of cycles. NARS is designed to operate incrementally, so longer or differently structured runs might produce additional results beyond what we captured.

Future Work

Future work can expand on this project in several directions. First, testing additional ethical dilemmas such as fairness versus utility or medical decision-making would help evaluate NARS's reasoning across different moral contexts. Second, experimenting with alternative Narsese encodings may lead to more refined or interpretable results. Finally, exploring a broader range of inputs, including temporal sequences, goal prioritization, or other inference patterns, could help assess the full potential of the system in modeling complex ethical reasoning.

Team Contributions

The project was completed by Manan Patel, Vivek Solanki, and Nathan Hollick. Contributions are outlined below:

- **Manan Patel:** Designed and encoded the problem in Narsese, conducted the evaluation using the NARS system, analyzed the system's outputs, and wrote the final report.
- **Vivek Solanki:** Participated in initial brainstorming, proposed a moral dilemma for evaluation, contributed to discussions about representing ethical scenarios in Narsese, and researched and finalized a suitable problem for evaluation with respect to NARS.
- **Nathan Hollick:** Participated in initial brainstorming, proposed a moral dilemma for evaluation, contributed to discussions about representing ethical scenarios in Narsese, and researched and finalized a suitable problem for evaluation with respect to NARS.

References

[1] John McCarthy et al. "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955". In: *AI magazine* 27.4 (1955), pp. 12–12.

[2] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[3] Daniel Crevier. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc., 1993.

[4] Tom M Mitchell and Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[6] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[7] Pei Wang. *Non-axiomatic logic: A model of intelligent reasoning*. World Scientific, 2013.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014).

[9] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[10] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[11] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[12] Pei Wang. *Axiomatic Reasoning in NARS*. Tech. rep. 2022.

[13] J OpenAI Achiam et al. "GPT-4 technical report. arXiv". In: *arXiv preprint arXiv:2303.08774* (2023).