Automating the Modelling of Transformative Artificial Intelligence Risks

An Epistemic Framework for Leveraging Frontier AI Systems to Upscale Conditional Policy Assessments in Bayesian Networks on a Narrow Path towards Existential Safety

Valentin Jakob Meyer Prof. Dr. Timo Speith

2025-05-26

The coordination crisis in AI governance presents a paradoxical challenge: unprecedented investment in AI safety coexists alongside fundamental coordination failures across technical, policy, and ethical domains. These divisions systematically increase existential risk by creating safety gaps, misallocating resources, and fostering inconsistent approaches to interdependent problems. This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach that addresses this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models.

The AMTAIR system implements an end-to-end pipeline that transforms unstructured text into interactive Bayesian networks through a novel two-stage extraction process: first capturing argument structure in ArgDown format, then enhancing it with probability information in BayesDown. This approach bridges communication gaps between stakeholders by making implicit models explicit, enabling comparison across different worldviews, providing a common language for discussing probabilistic relationships, and supporting policy evaluation across diverse scenarios.

```
# [ ]: Configure the Quarto Manuscript options: https://quarto.org/docs/manuscripts/componen
```

1

2 Frontmatter

3 Prefatory Apparatus: Illustrations and Terminology — Quick References

3.1 List of Tables

Table 1: Table name

Table 2: Table name

Table 3: Table name

3.2 List of Graphics & Figures

3.3 List of Abbreviations

esp. especially

f., ff. following

incl. including

p., pp. page(s)

MAD Mutually Assured Destruction

3.4 Glossary

4

5 Introduction

10% of Grade:

• introduces and motivates the core question or problem • provides context for discussion (places issue within a larger debate or sphere of relevance) • states precise thesis or position the author will argue for • provides roadmap indicating structure and key content points of the essay

~ 14% of text ~ 4200 words

- introduces and motivates the core question or problem

5.1 Motivation: Problem Statement

5.2 Motivation: Research Question

- provides context for discussion (places issue within a larger debate or sphere of relevance)

5.3 Scope: Aim & Context of the Research

5.4 Significance of the Research: Theory of Change

- states precise thesis or position the author will argue for

5.5 Thesis Statement & Position: (Aim of the Paper)

- provides roadmap indicating structure and key content points of the essay

5.6 Overview: Structure & Approach of the Paper (Roadmap — Theory of Change)

5.7 Table of Contents

6

7 Context

20% of Grade:

• demonstrates understanding of all relevant core concepts • explains why the question/thesis/problem is relevant in student's own words (supported by quotations) • situates it within the debate/course material • reconstructs selected arguments and identifies relevant assumptions • describes additional relevant material that has been consulted and integrates it with the course material as well as the research question/thesis/problem

~ 29% of text ~ 8700 words

1. successively (chunk my chunk) introduce concepts/ideas — and 2. ground each with existing literature

8

9 AMTAIR

20% of Grade:

• provides critical or constructive evaluation of positions introduced • develops strong (plausible) argument in support of author's own position/thesis • argument draws on relevant course material • claim/argument demonstrates understanding of the course materials incl. key arguments and core concepts within the debate • claim/argument is original or insightful, possibly even presents an original contribution to the debate

~ 29% of text ~ 8700 words

10

11 Discussion

10% of Grade:

• discusses a specific objection to student's own argument • provides a convincing reply that bolsters or refines the main argument • relates to or extends beyond materials/arguments covered in class

~ 14% of text ~ 4200 words

12

13 Conclusion

10% of Grade:

• summarizes thesis and line of argument • outlines possible implications • notes outstanding issues / limitations of discussion • points to avenues for further research • overall conclusion is in line with introduction

~ 14% of text ~ 4200 words

Bibliography/References

14

15 Appendices

16 Appendix A

17 Appendix B

18 Appendix C

19 Appendix D

TestText

20 Affidavit

21 Notebooks