UNIVERSITÄT BAYREUTH

– P&E Master's Programme –
Chair of Philosophy, Computer Science & Artificial Intelligence

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 Existencial Safety"

—

A thesis submitted at the Department of Philosophy

for the degree of *Master of Arts in Philosophy & Economics*

Author:

Valentin Jakob Meyer
Valentin.meyer@uni-bayreuth.de
Matriculation Number: 1828610
Tel.: +49 (1573) 4512494
Pielmühler Straße 15
52066 Lappersdorf

Supervisor:

Dr. Timo Speith

Word Count:
30.000
Source / Identifier:
Document URL

26th of May 2025

Abstract

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.

Source: Article Notebook

0.1 Grading

0.1.1 Research (10%)

- demonstrates knowledge of the subject area as drawn from appropriate sources
- incorporates insights from in-class discussions
- draws on appropriate additional materials beyond those covered in class (primary as well as secondary sources)
- covers relevant material at appropriate level of detail

0.2 Callout Test — Language & Style

• employs appropriate tone and academic language • uses effective and sophisticated sentence variety, diction, and vocabulary • employs correct English spelling and grammar • is clearly written and uses appropriate sentence complexity • communicates main points effectively / is easy to follow • formats citations and references correctly and consistently (e.g. (AUTHOR, YEAR) citation style) • names all primary and secondary sources • includes a complete list of references with full bibliographic details

More text

Chapter 1

Introduction

1.1 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

1.2 Motivation: Problem Statement

1.3 Motivation: Research Question

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

1.4 Scope: Aim & Context of the Research

1.5 Significance of the Research: Theory of Change

• states precise thesis or position the author will argue for

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

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

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

1.8 Table of Contents

Source: Introduction

Chapter 2

Context

2.0.1 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

2.1 Theoretical Background Considerations

2.1.1 DAG / BayesNets

2.1.2 State of the art (MTAIR) — Explanation

Carlsmith Model (Analytica)

2.1.3 (Intro) Example — Rain/Sprinkler/Lawn

/ Rain/Sprinkler/Lawn DAG / BayesNet — Extended Example

…

Own Position/Argument: AMTAIR … Own Rain/Sprinkler/Lawn DAG / BayesNet Implementation

2.2 Methodology

MTAIR / Carlsmith Model (Analytica) — Explanation (— is motivation: should come first)

2.2.1 Kialo

2.2.2 Rain/Sprinkler/Lawn DAG

2.2.3 BayeServer

2.2.4 BayesNet — Extended Example

2.2.5 Code + documentation

Source: Context

Chapter 3

AMTAIR

3.1 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

Chapter 4

Implementation

TestText

Chapter 5

Results

TestText

5.1 Own Carlsmith Model Implementation — Explanation

5.2 Own Implementation: Good example from a published paper

Source: AMTAIR

Chapter 6

Discussion

6.1 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

Source: Discussion

Chapter 7

Conclusion

7.1 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

Source: Conclusion

Bibliography/References

Prefatory Apparatus: Illustrations and Terminology — Quick References

List of Tables

List of Graphics & Figures

List of Abbreviations

esp. especially
f., ff. following
incl. including
p., pp. page(s)
MAD Mutually Assured Destruction

Glossary

term Definition of term

Another term Description of second term

Text

Chapter 8

Appendices

8.1 Appendices

8.2 Appendix A

8.3 Appendix B

8.4 Appendix C

8.5 Appendix D

TestText

Source: Appendices

Notebooks

UNIVERSITÄT BAYREUTH

– P&E Master's Programme –
Chair of Philosophy, Computer Science & Artificial Intelligence

Affidavit

Declaration of Academic Honesty

Hereby, I attest that I have composed and written the presented thesis

Automating the Modelling of Transformative Artificial Intelligence Risks

independently on my own, without the use of other than the stated aids and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

This paper has neither been previously submitted in the same or a similar form to another authority nor has it been published yet.

Bayreuth on the
May 12, 2025

Valentin Meyer