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 Structure (10%)

- outlines project clearly in the introduction
- discussion follows a logical order
- order of sections corresponds to outline
- uses appropriate topic and transition sentences
- employs proper signposting and cross-referencing throughout paper
- sections are appropriately numbered and headlined

0.3 Callout Test — Language & Style (10%)

- 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

1.9 Problem Statement — Motivation

Continued AI Progress:

- Rapid advancements in AI technology increase both potential benefits and risks.

Existential Risks (AI X-Risk):

- Advanced AI systems could pose significant threats if misaligned with human values.

Complexity Challenges:

- The intricate nature of AI systems complicates policy formulation and understanding.

Limitations of Current Approaches:

- MTAIR's Reliance on Human Labor:
  - Modeling Transformative AI Risks (MTAIR) is constrained by manual cognitive efforts.

- Need for Automation:
  - Scaling and automating risk modeling is essential to keep pace with AI developments.

Opportunity:

- Leveraging new technologies to enhance our ability to model and mitigate AI risks.

1.10 Aim of the Paper

1.10.1 Research Question & Scope

Can frontier AI technologies be utilized to automate the modeling of transformative AI risks, so as to allow for the prediction of policy impacts?

Frontier AI Technology: Today's most capable AI systems (e.g. GPT4 level LLMs)
Scaling Up: Automating the previously "manual" cognitive labor
Modeling: Formalizing the world views underlying arguments
Transformative AI: Level of AI capabilities defined by severe impact on the world
Safety & Governance Literature: Publications, reports etc. concerned with risks from AI
Automated Estimation: Non-manual (AI systems + scaffolding), quantified evaluations
Probability Distributions: Formal expressions of the expectations over future worlds
Conditional Trees of Possible Worlds: "If … then…" reasoning over ways things may play out
Forecasting Policy Impacts: Qualitative & quantitative evaluation of expected outcomes

1.10.2 Significance of the Research

1.10.3

1.11 Theory of Change — Approach & Structure of the Paper

Multiplicative Benefits:

- Automation × Live Prediction Market Integrations × Policy Impact Evaluations

Explanation:
Automation:

- Increases efficiency and scalability of risk modeling.

  Live Prediction Markets:

- Provides up-to-date, collective intelligence to inform models.

  Policy Impact Evaluations:

- Improves the accuracy and relevance of policy assessments.

Outcome:

- Enhanced ability to develop effective policies that mitigate AI risks.

Visual Aid:

- A diagram illustrating how each component amplifies the others, leading to greater overall impact.

1.12

1.13 Overview / 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

3.0.1 20% of Grade: ~ 29% of text ~ 8700 words

- 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
- demonstrate 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

3.1 Own Carlsmith Model Implementation — Explanation

3.2 Own Implementation: Good example from a published paper

3.3 Implementation

TestText

3.4 Results

TestText

Source: AMTAIR

Chapter 4

Discussion

4.1 Discussion

10% of Grade: ~ 14% of text ~ 4200 words

- 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

Chapter 5

Discussion — Exchange, Controversy & Influence

5.1 Challenges & Problems — Red Teaming Problems, Failures & Downsides

`Potential Failures:`

- Data Issues: Inaccurate or biased inputs.
- Model Limitations: Oversimplifications.
- Tech Risks: AI misinterpretations.
  Red Teaming:
- Stress-testing models to find weaknesses.
  Impact on Theory of Change:
- Identifying points of failure strengthens the approach.

5.2 Implications & Impact — Uptake, Feedback Loops, Uptake & Success – Green Teaming –

`Potential Outcomes:`

- First-Order: Reduced AI risks through better policies.
- Second-Order: Enhanced collaboration.
- Third-Order: Framework applied to other global risks.
  Feedback Loops:
- Continuous model improvement.
- Adaptive policy-making.
  Green Teaming:
- Strategies to maximize positive impacts.

5.3 Known Unknowns & Unknown Unknowns — Input Data Example: Modeling Author Worldviews from Bibliographies Instead of Individual Papers

`Potential Outcomes:`

- First-Order: Reduced AI risks through better policies.
- Second-Order: Enhanced collaboration.
- Third-Order: Framework applied to other global risks.
  Feedback Loops:
- Continuous model improvement.
- Adaptive policy-making.
  Green Teaming:
- Strategies to maximize positive impacts.

Source: Discussion

Chapter 6

Conclusion

6.1 The Current State of Things & How to Continue

10% of Grade: ~ 14% of text ~ 4200 words

- 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

6.2 Summary — Key Takeaways & Findings

6.2.1 Assessing Policy Effects:

Evaluating how different policies alter P(Doom).

6.2.2 Conditional Probability:

Calculating P(Doom | Policy Alpha).

6.2.3 Methodology:

Update model parameters based on policy implementation.
Recompute probabilities accordingly.

6.2.4 Purpose:

Inform policymakers of potential policy effectiveness.
Prioritize interventions that significantly reduce risks.

6.3 Outlook — Outlook & Next Steps / Further Research

6.3.1 Scaling Up:

- Include more variables and data sources.

6.3.2 Collaboration:

- Partner with policymakers and researchers.

6.3.3 Technological Enhancements:

- Employ advanced AI techniques.

6.3.4 Potential Impact:

- Influence global AI governance.

6.3.5 Limitations of the Analysis

6.3.6 Policy Implications & Recommendations

6.3.7 Areas for Future Research

6.3.8 Open Questions — Central/Remaining Questions & Feedback

Questions:

- How can we improve automation accuracy?
- What challenges exist in policy implementation?
- How do we mitigate AI model biases?
- How can interdisciplinary efforts enhance outcomes?

Feedback:

- Invite thoughts, critiques, and suggestions.

6.3.9 Outlook — Outlook & Next Steps / Further Research

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 7

Appendices

7.1 Appendices

7.2 Appendix A

7.3 Appendix B

7.4 Appendix C

7.5 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 15, 2025

Valentin Meyer