Automating the Modelling of Transformative Artificial Intelligence Risks

Valentin Jakob Meyer

2025-05-26

Table of contents

# Preface

This is a Quarto book.

To learn more about Quarto books visit <https://quarto.org/docs/books>.

# Abstract

# Outline(s): Table of Contents

# 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

|  |
| --- |
| Figure 1.1: Caption/Title 2 |

Testing crossreferencing grapics [Figure 1.1](#fig-testgraphic2).

## 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

#### 1.10.1.1 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

## 1.11 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

# 2. 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

## 2.1 Theoretical Background Considerations

## 2.2 Literature, Concepts & Terminology

### 2.2.1 DAG / BayesNets

### 2.2.2 State of the art (MTAIR) — Explanation

#### 2.2.2.1 Carlsmith Model (Analytica)

### 2.2.3 (Intro) Example — Rain/Sprinkler/Lawn

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

…

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

## 2.3 Methodology

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

### 2.3.1 Kialo

### 2.3.2 Rain/Sprinkler/Lawn DAG

### 2.3.3 BayeServer

### 2.3.4 BayesNet — Extended Example

### 2.3.5 Code + documentation

|  |
| --- |
| [AMTAIR Automation Pipeline](https://github.com/VJMeyer/submission)  Figure 2.1: AMTAIR Automation Pipeline |

Testing crossreferencing grapics [Figure 2.1](#fig-automation_pipeline).

# 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

# 4. Insights & Findings

## 4.1 Automated Modeling Pipeline — From Academic Papers to Political Strategy

Success of Automation:

* Demonstrated feasibility of automated model extraction.
* Improved Forecasting:
* Enhanced accuracy with real-time data integration.
* Policy Analysis:
* Identified impactful policies through conditional forecasting.
* Scalability Achieved:
* Efficient processing of extensive data sets.
* Addressed Challenges:
* Overcame limitations of manual modeling.

## 4.2 Project Scaling — Workflow Pipeline & Automation

Scaling Opportunities:

* Horizontal: Incorporate more data sources.
* Vertical: Add detailed variables.

New Capabilities:

* Advanced analytics.
* Real-time data integration.

Requirements:

* Software Setup: Robust infrastructure.
* Financial: Funding for APIs and compute resources.

Impact:

* Broader, more comprehensive models.
* Enhanced policy analysis.

## 4.3 Computational Complexity — Computational Tractability

Challenges:

* High computational demands of complex models.

Solutions:

* Clustering Worldviews:
* Group similar perspectives to simplify models.
* Correlation Management:
* Adjust for variable interdependencies.
* Efficient Algorithms:  
  Use optimized sampling methods like Monte Carlo.

Outcome:

* Achieved efficiency without sacrificing accuracy.

Link to Theory of Change:

* Scalability amplifies policy impact.

## 4.4 External Validation — Manual Extraction & Processing

Purpose:

* Assess accuracy of automated methods.
* Comparison:
* Automation Strengths:
* Speed, consistency.
* Human Strengths:
* Nuanced understanding.
* Findings:
* Automation excels in data handling.
* Human oversight enhances quality.
* Conclusion:
* Optimal results from combining AI with expert input.

# 5. Discussion

## 5.1 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

# 6. Discussion — Exchange, Controversy & Influence

## 6.1 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.

## 6.2 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.

## 6.3 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.

# 7. Conclusion

## 7.1 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

## 7.2 Summary — Key Takeaways & Findings

### 7.2.1 Assessing Policy Effects:

Evaluating how different policies alter P(Doom).

### 7.2.2 Conditional Probability:

Calculating P(Doom | Policy Alpha).

### 7.2.3 Methodology:

Update model parameters based on policy implementation.

Recompute probabilities accordingly.

### 7.2.4 Purpose:

Inform policymakers of potential policy effectiveness.

Prioritize interventions that significantly reduce risks.

## 7.3 Outlook — Outlook & Next Steps / Further Research

### 7.3.1 Scaling Up:

* Include more variables and data sources.

### 7.3.2 Collaboration:

* Partner with policymakers and researchers.

### 7.3.3 Technological Enhancements:

* Employ advanced AI techniques.

### 7.3.4 Potential Impact:

* Influence global AI governance.

### 7.3.5 Limitations of the Analysis

### 7.3.6 Policy Implications & Recommendations

### 7.3.7 Areas for Future Research

### 7.3.8 Open Questions — Central/Remaining Questions & Feedback

#### 7.3.8.1 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?

#### 7.3.8.2 Feedback:

* Invite thoughts, critiques, and suggestions.

### 7.3.9 Outlook — Outlook & Next Steps / Further Research

# Frontmatter

# Prefatory Apparatus: Illustrations and Terminology — Quick References

## List of Tables

Table 1: Table name

Table 2: Table name

Table 3: Table name

* Figure 1.1: The coordination crisis in AI governance - visualization of fragmentation
* Figure 2.1: The Carlsmith model - DAG representation
* Figure 3.1: Research design overview - workflow diagram
* Figure 3.2: From natural language to BayesDown - transformation process
* Figure 4.1: ARPA system architecture - component diagram
* Figure 4.2: Visualization of Rain-Sprinkler-Grass\_Wet Bayesian network - screenshot
* Figure 5.1: Extraction quality metrics - comparative chart
* Figure 5.2: Comparative analysis of AI governance worldviews - network visualization
* Table 2.1: Comparison of approaches to AI risk modeling
* Table 3.1: Probabilistic translation guide for qualitative expressions
* Table 4.1: System component responsibilities and interactions
* Table 5.1: Policy impact evaluation results - summary metrics

## List of Graphics & Figures

## List of Abbreviations

esp. especially

f., ff. following

incl. including

p., pp. page(s)

MAD Mutually Assured Destruction

* AI - Artificial Intelligence
* AGI - Artificial General Intelligence
* ARPA - AI Risk Pathway Analyzer
* DAG - Directed Acyclic Graph
* LLM - Large Language Model
* MTAIR - Modeling Transformative AI Risks
* P(Doom) - Probability of existential catastrophe from misaligned AI
* CPT - Conditional Probability Table

## Glossary

* **Argument mapping**: A method for visually representing the structure of arguments
* **BayesDown**: An extension of ArgDown that incorporates probabilistic information
* **Bayesian network**: A probabilistic graphical model representing variables and their dependencies
* **Conditional probability**: The probability of an event given that another event has occurred
* **Directed Acyclic Graph (DAG)**: A graph with directed edges and no cycles
* **Existential risk**: Risk of permanent curtailment of humanity’s potential
* **Power-seeking AI**: AI systems with instrumental incentives to acquire resources and power
* **Prediction market**: A market where participants trade contracts that resolve based on future events
* **d-separation**: A criterion for identifying conditional independence relationships in Bayesian networks
* **Monte Carlo sampling**: A computational technique using random sampling to obtain numerical results

# Headings & Potential Headings

Notes

Ideas

Citations / Things to reference

verbatim code

Blockquote

Here is an inline note.[[1]](#footnote-106)

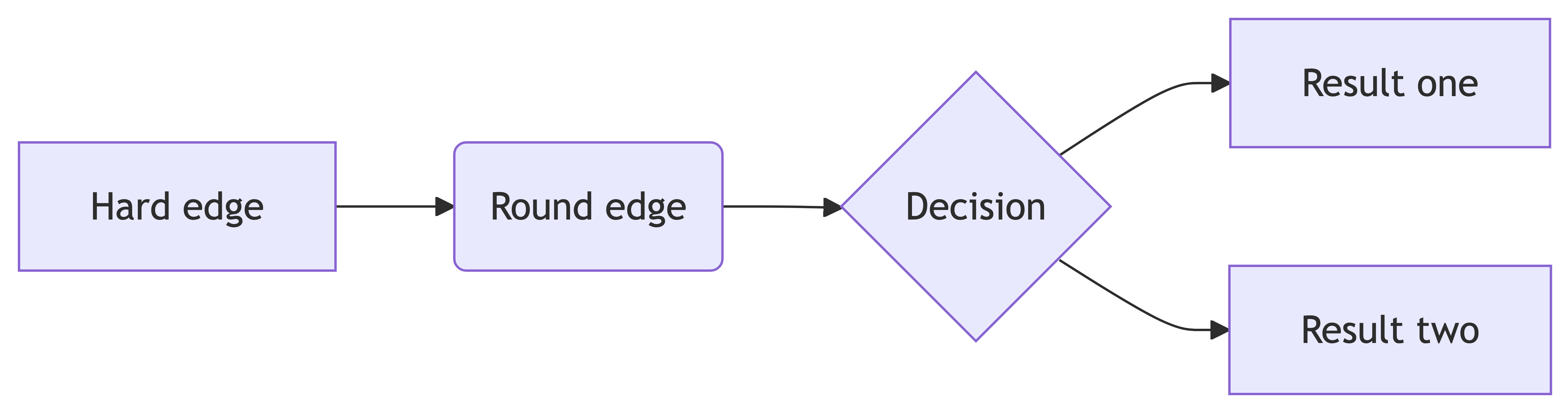
Here is a footnote reference,[[2]](#footnote-107)

code

Here’s some raw inline HTML:

page 1

page 2



Testing crossreferencing grapics [Figure 2.1](#fig-automation_pipeline).

Referencing Sections **?@sec-citations**.

headings as **?@sec-heading**.

## Citations

(Knuth 1984)

Adams (1996)

(Adams 1996)

### Quarto Features Incompatible with LaTeX (Below)

# Bibliography (References)

Adams, Ernest W. 1996. “Four Probability-Preserving Properties of Inferences.” *Journal of Philosophical Logic* 25: 1–24.

Knuth, Donald E. 1984. “Literate Programming.” *Computer Journal* 27 (2): 97–111. <https://doi.org/10.1093/comjnl/27.2.97>.

# Appendix A — Appendices

## A.1 Appendices

## A.2 Appendix A

## A.3 Appendix B

## A.4 Appendix C

## A.5 Appendix D

TestText

# Appendix B — appendixA

testtext

1. Inlines notes are easier to write, since you don’t have to pick an identifier and move down to type the note. [↑](#footnote-ref-106)
2. Here is the footnote. [↑](#footnote-ref-107)