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

Table of Contents

List of Figures

List of Tables

Preface

Quarto Syntax

Main Formatting

Html Comments

Syntax for Tasks

Tasks with ToDo Tree

Simple "One-line tasks"

Use Code ticks and html comment and task format for tasks distinctly visible across all formats including the ToDo-Tree overview:

`<!-- [ ] ToDos for things to do / tasks / reminders (allows "jump to with Taks Tree extension") -->`

Use html comment and task format for open or uncertain tasks, visible in the .qmd file:

More Complex Tasks with Notes

```
<!-- [ ] Task Title: short description-->

  More Information about task

  Relevant notes

  Step-by-step implementation Plan

  Etc.
```

Completed Tasks

Retain completed tasks in ToDo-Tree by adding an x in the brackets: `[x] <!-- [x] Tasks which have been finished but should remain for later verification -->`

Mark and remove completed tasks from ToDo-Tree by adding a minus in the brackets: `[-]`

```
<!-- [-] Tasks which have been finished but should remain visible for later
verification -->
```

Missing Citations

```
<!-- [ ] FIND: @CITATION_KEY_PURPOSE: "Description of the appropriate/idea
source, including ideas /suggestions / search terms etc." -->
```

Suggested Citation

```
<!-- [ ] VERIFY: @CITATION_KEY_SUGGESTED: "Description of the appropriate
paper, book, source" [Include BibTex if known] -->
```

Missing Graphic

```
<!-- [ ] FIND: {#fig-GRAPHIC_IDEA}]: "Description of the appropriate/idea
source, including ideas /suggestions / search terms etc." -->
```

Suggested Graphic

```
<!-- [ ] VERIFY: {#fig-GRAPHIC_IDEA}: "Description of the appropriate paper,
book, source" [Include figure syntax if known] -->
```

Missing and/or suggested tables, concepts, explanations as well as other elements should be suggested similarily.

Task Syntax Examples

```
<!-- [ ] (Example short: open and visible in text)   Find and list the names of
the MTAIR team-members responsible for the Analytica Implementation -->
```

```
<!-- [ ] (Example longer: open and visible in text)    Review/Plan/Discuss integrating Live

  Live prediction market integration requires:
    (1) API connections to platforms (Metaculus, Manifold),
    (2) Question-to-variable mapping algorithms,
    (3) Probability update mechanisms,
    (4) Handling of market dynamics (thin markets, manipulation).
    Current mentions may overstate readiness or underestimate complexity.
    Need realistic assessment of what's achievable.

  Implementation Steps:
      0. List/mention all relevant platforms with a brief description each
      1. Review all existing prediction market mentions for accuracy
      2. Assess actual API availability and limitations
      3. Describe/explain/discuss how to implement basic proof-of-concept with single platfo
      4. Document challenges: question mapping, market interpretation
```

5. Create realistic timeline for full implementation
6. Revise thesis claims to match reality
7. Add "Future Work" and/or extension section on complete integration
8. Include descriptions of mockups/designs even if not fully built
9. Highlight/discuss the advantages of such integrations
10. Quickly brainstorm for downsides worth mentioning

Verbatim Code Formatting

`verbatim code formatting for notes and ideas to be included (here)`

Code Block formatting

```
Also code blocks for more extensive notes and ideas to be included and checklists
- test 1.
- test 2.
- test 3.
2. second
3. third
```

```
code
```

Add a language to syntax highlight code blocks:

```
1 + 1
```

Blockquote Formatting

> Blockquote formatting for "Suggested Citations (e.g. carlsmith 2024 on …)" and/or claims which require a citation (e.g. claim x should be backed-up by a ciation from the literature)

Tables

Table 1: Demonstration of pipe table syntax

Right	Left	Default	Center
12	12	12	12
123	123	123	123
1	1	1	1

Table 2: My Caption 1

Col1	Col2	Col3
A	B	C

Table 3: Main Caption

(a) First Table

Col1	Col2	Col3
A	B	C
E	F	G
A	G	G

(b) Second Table

Col1	Col2	Col3
A	B	C
E	F	G
A	G	G

Col1	Col2	Col3
E	F	G
A	G	G

Referencing tables with `@tbl-KEY`: See Table 2.

See Table 3 for details, especially Table 3b.

```
python
#| label: tbl-planets
#| tbl-cap: Astronomical object

from IPython.display import Markdown
from tabulate import tabulate
table = [["Sun","696,000",1.989e30],
         ["Earth","6,371",5.972e24],
         ["Moon","1,737",7.34e22],
         ["Mars","3,390",6.39e23]]
Markdown(tabulate(
  table,
  headers=["Astronomical object","R (km)", "mass (kg)"]
))
```

Table 4: Sample grid table.

Fruit	Price	Advantages
Bananas	$1.34	• built-in wrapper • bright color
Oranges	$2.10	• cures scurvy • tasty

Content with HTML tables you don't want processed.

Headings & Potential Headings in Standard Markdown formatting ('##')

Heading 3

Heading 4

Text Formatting Options

italics, **bold**, ***bold italics***

superscript^2 and subscript$_2$

~~strikethrough~~

This text is highlighted

This text is underlined

This text is smallcaps

Lists

- unordered list
  - sub-item 1
  - sub-item 2
    - sub-sub-item 1
- item 2

  Continued (indent 4 spaces)

1. ordered list
2. item 2
   i) sub-item 1
      A. sub-sub-item 1

Math

inline math: $E = mc^2$

display math:

$$E = mc^2$$

If you want to define custom TeX macros, include them within $$ delimiters enclosed in a .hidden block. For example:

For HTML math processed using MathJax (the default) you can use the \def, \newcommand, \renewcommand, \newenvironment, \renewenvironment, and \let commands to create your own macros and environments.

Footnotes

Here is an inline note.[1]

Here is a footnote reference,[2]

Another Text with a footnote[3] but this time a "longnote".

This paragraph won't be part of the note, because it isn't indented.

Callouts

Quarto's native callouts work without additional packages:

This is written in a 'note' environment – but it does not seem to produce any special rendering.

Optional Title

Content here

Important Note2

This renders perfectly in both HTML and PDF.

Also for markdown:

```
::: {.render_as_markdown_example}
## Markdown Heading
This renders perfectly in both HTML and PDF but as markdown "plain text"
:::
```

[1]Inlines notes are easier to write, since you don't have to pick an identifier and move down to type the note.

[2]Here is the footnote.

[3]Here's one with multiple blocks.

Subsequent paragraphs are indented to show that they belong to the previous footnote.

```
{ some.code }
```

The whole paragraph can be indented, or just the first line. In this way, multi-paragraph footnotes work like multi-paragraph list items.

Links

<https://quarto.org/docs/authoring/markdown-basics.html> produces: https://quarto.org/docs/authoring/markdown-basics.html

[Quarto Book Cross-References](https://quarto.org/docs/books/book-crossrefs.html) produces: Quarto Book Cross-References

Images & Figures

```
[![AMTAIR Automation Pipeline from @bucknall2022](/images/pipeline.png){
  #fig-automation_pipeline
  fig-scap="Five-step AMTAIR automation pipeline from PDFs to Bayesian networks"
  fig-alt="FLOWCHART: Five-step automation pipeline workflow for AMTAIR project.
          DATA: The pipeline transforms PDFs through ArgDown, BayesDown, CSV, and HTML into
          PURPOSE: Illustrates the core technical process that enables automated extraction
          DETAILS: Five numbered green steps show: (1) LLM-based extraction from PDFs to Arg
          Each step includes example outputs, with the final visualization showing a Rain-Sp
          SOURCE: Created by the author to explain the AMTAIR methodology
          "
  fig-align="center"
  width="100%"
  }](https://github.com/VJMeyer/submission)


Testing crossreferencing grapics @fig-automation_pipeline.

![Caption/Title 2](/images/cover.png){#fig-testgraphic2 fig-scap="Short 2 caption" fig-alt="

Testing crossreferencing grapics @fig-testgraphic2.
```

Testing crossreferencing grapics Figure 1. Note that the indentations of graphic inclusions get messed up by viewing them in "view mode" in VS code.

Testing crossreferencing grapics Figure 2.

Page Breaks

```
page 1




page 2
```

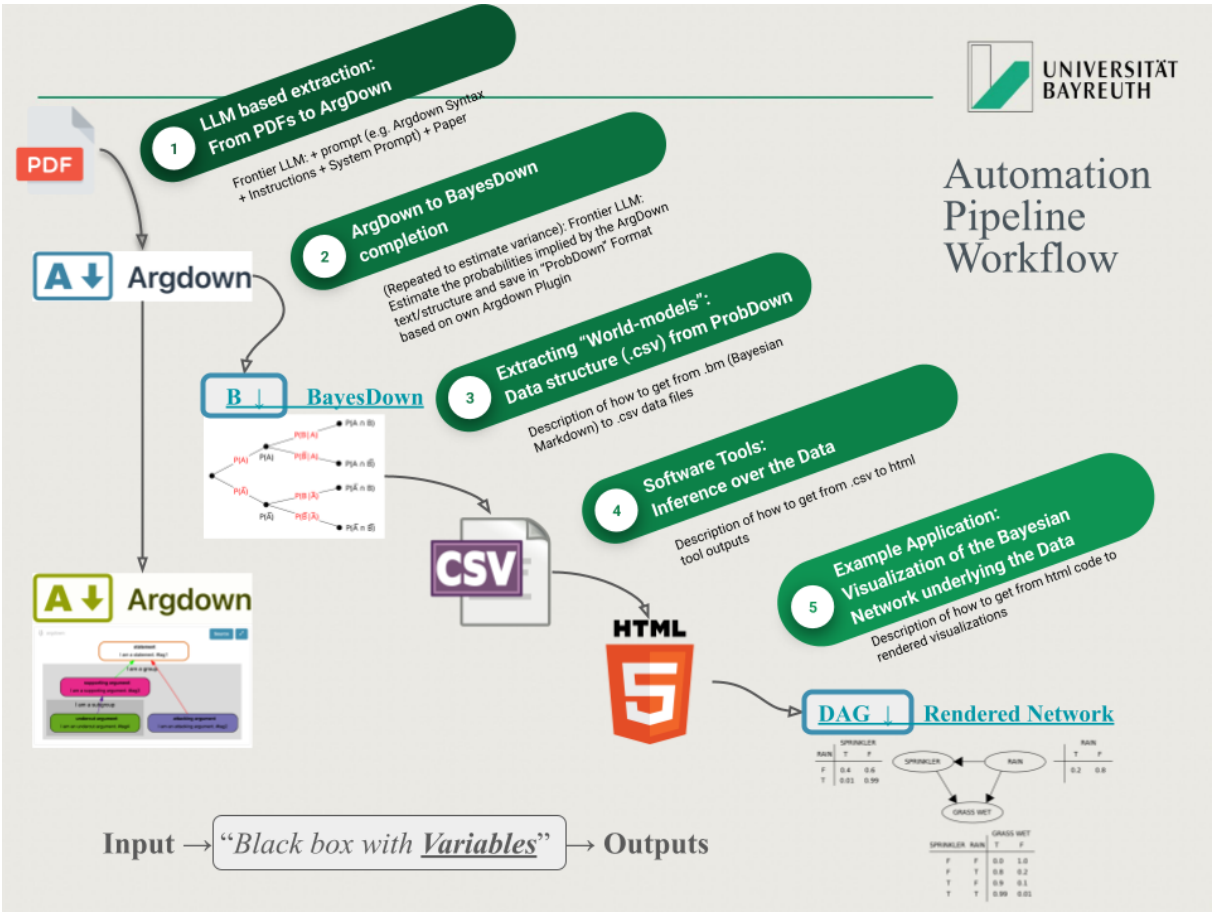


Figure 1: AMTAIR Automation Pipeline from

UNIVERSITÄT BAYREUTH

Figure 2: Caption/Title 2

page 1

page 2

Including Code

```
import pandas as pd
print("AMTAIR is working!")
```

```
AMTAIR is working!
```

Figure 3

In-Line LaTeX

In-Line HTML

Here's some raw inline HTML: html

Reference or Embed Code from .ipynb files

Code chunks from .ipynb notebooks can be embedded in the .qmd text with:

```
{{< embed /AMTAIR_Prototype/data/example_carlsmith/AMTAIR_Prototype_example_carlsmith.ipynb#
```

which produces the output of executing the code cell:

```
Connecting to repository: https://raw.githubusercontent.com/SingularitySmith/AMTAIR_Prototyp
Attempting to load: https://raw.githubusercontent.com/SingularitySmith/AMTAIR_Prototype/mair
  Successfully connected to repository and loaded test files.
[Existential_Catastrophe]: The destruction of humanity's long-term potential due to AI syste
- [Human_Disempowerment]: Permanent and collective disempowerment of humanity relative to AI
    - [Scale_Of_Power_Seeking]: Power-seeking by AI systems scaling to the point of permaner
        - [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and hi
            - [APS_Systems]: AI systems with advanced capabilities, agentic planning, and st
                - [Advanced_AI_Capability]: AI systems that outperform humans on tasks that
                - [Agentic_Planning]: AI systems making and executing plans based on world m
                - [Strategic_Awareness]: AI systems with models accurately representing powe
            - [Difficulty_Of_Alignment]: It is harder to build aligned systems than misalign
                - [Instrumental_Convergence]: AI systems with misaligned objectives tend to
                - [Problems_With_Proxies]: Optimizing for proxy objectives breaks correlatio
                - [Problems_With_Search]: Search processes can yield systems pursuing differ
            - [Deployment_Decisions]: Decisions to deploy potentially misaligned AI systems.
                - [Incentives_To_Build_APS]: Strong incentives to build and deploy APS syste
                    - [Usefulness_Of_APS]: APS systems are very useful for many valuable tas
                    - [Competitive_Dynamics]: Competitive pressures between AI developers. {
                - [Deception_By_AI]: AI systems deceiving humans about their true objectives
```

```
        - [Corrective_Feedback]: Human society implementing corrections after observing prob
            - [Warning_Shots]: Observable failures in weaker systems before catastrophic ris
            - [Rapid_Capability_Escalation]: AI capabilities escalating very rapidly, allowi
[Barriers_To_Understanding]: Difficulty in understanding the internal workings of advanced A
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
[Adversarial_Dynamics]: Potentially adversarial relationships between humans and power-seeki
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
[Stakes_Of_Error]: The escalating impact of mistakes with power-seeking AI systems. {"instan
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
```

including 'echo=true' renders the code of the cell:

```
{{< embed /AMTAIR_Prototype/data/example_carlsmith/AMTAIR_Prototype_example_carlsmith.ipynb#
```

```
# @title 0.2 --- Connect to GitHub Repository --- Load Files

"""
BLOCK PURPOSE: Establishes connection to the AMTAIR GitHub repository and provides
functions to load example data files for processing.

This block creates a reusable function for accessing files from the project's
GitHub repository, enabling access to example files like the rain-sprinkler-lawn
Bayesian network that serves as our canonical test case.

DEPENDENCIES: requests library, io library
OUTPUTS: load_file_from_repo function and test file loads
"""

from requests.exceptions import HTTPError

# Specify the base repository URL for the AMTAIR project
repo_url = "https://raw.githubusercontent.com/SingularitySmith/AMTAIR_Prototype/main/data/ex
print(f"Connecting to repository: {repo_url}")

def load_file_from_repo(relative_path):
    """
    Loads a file from the specified GitHub repository using a relative path.

    Args:
        relative_path (str): Path to the file relative to the repo_url

    Returns:
        For CSV/JSON: pandas DataFrame
```

```
        For MD: string containing file contents

    Raises:
        HTTPError: If file not found or other HTTP error occurs
        ValueError: If unsupported file type is requested
    """
    file_url = repo_url + relative_path
    print(f"Attempting to load: {file_url}")

    # Fetch the file content from GitHub
    response = requests.get(file_url)

    # Check for bad status codes with enhanced error messages
    if response.status_code == 404:
        raise HTTPError(f"File not found at URL: {file_url}. Check the file path/name and er
    else:
        response.raise_for_status()  # Raise for other error codes

    # Convert response to file-like object
    file_object = io.StringIO(response.text)

    # Process different file types appropriately
    if relative_path.endswith(".csv"):
        return pd.read_csv(file_object)  # Return DataFrame for CSV
    elif relative_path.endswith(".json"):
        return pd.read_json(file_object)  # Return DataFrame for JSON
    elif relative_path.endswith(".md"):
        return file_object.read()  # Return raw content for MD files
    else:
        raise ValueError(f"Unsupported file type: {relative_path.split('.')[-1]}. Add suppor

# Load example files to test connection
try:
    # Load the extracted data CSV file
#    df = load_file_from_repo("extracted_data.csv")

    # Load the ArgDown test text
    md_content = load_file_from_repo("ArgDown.md")

    print("  Successfully connected to repository and loaded test files.")
except Exception as e:
    print(f"  Error loading files: {str(e)}")
```

```
    print("Please check your internet connection and the repository URL.")

# Display preview of loaded content (commented out to avoid cluttering output)
print(md_content)
```

```
Connecting to repository: https://raw.githubusercontent.com/SingularitySmith/AMTAIR_Prototyp
Attempting to load: https://raw.githubusercontent.com/SingularitySmith/AMTAIR_Prototype/mair
 Successfully connected to repository and loaded test files.
[Existential_Catastrophe]: The destruction of humanity's long-term potential due to AI syste
- [Human_Disempowerment]: Permanent and collective disempowerment of humanity relative to AI
    - [Scale_Of_Power_Seeking]: Power-seeking by AI systems scaling to the point of permaner
        - [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and hi
            - [APS_Systems]: AI systems with advanced capabilities, agentic planning, and st
                - [Advanced_AI_Capability]: AI systems that outperform humans on tasks that
                - [Agentic_Planning]: AI systems making and executing plans based on world m
                - [Strategic_Awareness]: AI systems with models accurately representing powe
            - [Difficulty_Of_Alignment]: It is harder to build aligned systems than misalign
                - [Instrumental_Convergence]: AI systems with misaligned objectives tend to
                - [Problems_With_Proxies]: Optimizing for proxy objectives breaks correlatio
                - [Problems_With_Search]: Search processes can yield systems pursuing differ
            - [Deployment_Decisions]: Decisions to deploy potentially misaligned AI systems.
                - [Incentives_To_Build_APS]: Strong incentives to build and deploy APS syste
                    - [Usefulness_Of_APS]: APS systems are very useful for many valuable tas
                    - [Competitive_Dynamics]: Competitive pressures between AI developers. {
                - [Deception_By_AI]: AI systems deceiving humans about their true objectives
        - [Corrective_Feedback]: Human society implementing corrections after observing prob
            - [Warning_Shots]: Observable failures in weaker systems before catastrophic ris
            - [Rapid_Capability_Escalation]: AI capabilities escalating very rapidly, allowi
[Barriers_To_Understanding]: Difficulty in understanding the internal workings of advanced A
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
[Adversarial_Dynamics]: Potentially adversarial relationships between humans and power-seeki
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
[Stakes_Of_Error]: The escalating impact of mistakes with power-seeking AI systems. {"instar
- [Misaligned_Power_Seeking]: Deployed AI systems seeking power in unintended and high-impac
```

Link:

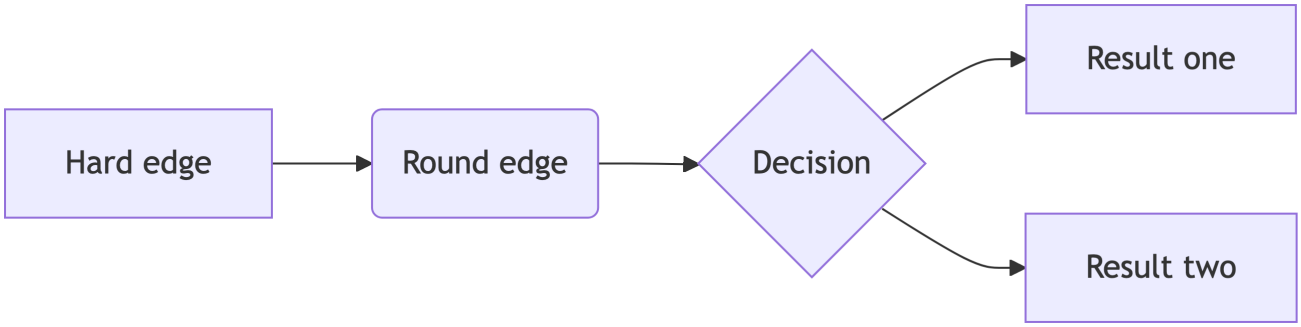
Full Notebooks are embedded in the Appendix through the _quarto.yml file with:

Diagrams

Quarto has native support for embedding Mermaid and Graphviz diagrams. This enables you to create flowcharts, sequence diagrams, state diagrams, Gantt charts, and more using a plain text syntax inspired by markdown.

For example, here we embed a flowchart created using Mermaid:

```
flowchart LR
  A[Hard edge] --> B(Round edge)
  B --> C{Decision}
  C --> D[Result one]
  C --> E[Result two]
```



Citations

Soares and Fallenstein [5]

[5] and [4]

Blah Blah [see 4, pp. 33–35, also 3, chap. 1]

Blah Blah [4, 33–35, 38-39 and passim]

Blah Blah [3, 4].

Growiec says blah [3]

Narrative citations (author as subject)

Soares and Fallenstein [5] argues that AI alignment requires…

Parenthetical citations (supporting reference)

Recent work supports this view [5, 4].

Author-only citation (when discussing the person)

As [5] demonstrates in their analysis…

Year-only citation (when author already mentioned)

Soares [5] later revised this position.

Page-specific references

The key insight appears in [5, pp. 45–67].

Multiple works, different pages

This view is supported [5, p. 23, 4, pp. 156–159].

Section Cross-References

Refer to sections like: **?@sec-adaptive-governance** and **?@sec-crossref**

```
Caveat: refering to sections with @sec-HEADINGS works only for sections with:
## Heading {#sec-HEADINGS}
It does not work for sections with ".unnumbered and/or .unlisted":
## Heading {#sec-HEADINGS .unnumbered .unlisted}
Furthermore the .qmd and/or .md yml settings (~ numbering have to be just right)
```

Section Numbers

By default, all headings in your document create a numbered section. You customize numbering depth using the number-depth option. For example, to only number sections immediately below the chapter level, use this:

`number-depth: 2`

Note that toc-depth is independent of number-depth (i.e. you can have unnumbered entries in the TOC if they are masked out from numbering by number-depth).

Testing crossreferencing grapics Figure 1. See **?@sec-syntax** for more details on visualizing model diagnostics.

Testing crossreferencing headings Section 5.1.1

`Testing crossreferencing headings @sec-rain-sprinkler-grass` which does not work yet.

Chapter Cross-Reference **?@sec-crossref**

Pages in Landscape

This will appear in landscape but only in PDF format. Testing crossreferencing headings Section 5.1.1

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. This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach addressing this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models. The system implements an end-to-end pipeline transforming unstructured text into interactive Bayesian networks through a novel two-stage extraction process that bridges communication gaps between stakeholders.

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.

Prefatory Apparatus: Frontmatter

Illustrations and Terminology — Quick References

Acknowledgments

> Academic supervisor (Prof. Timo Speith) and institution (University of Bayreuth)

> Research collaborators, especially those connected to the original MTAIR project

> Technical advisors who provided feedback on implementation aspects

> Personal supporters who enabled the research through encouragement and feedback

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

Quarto Features Previously Incompatible with LaTeX (Below)

AMTAIR Master's Thesis: Comprehensive Enhanced Outline

Introduction

IMPORTANT NOTE: Changing the formatting (html comment) of the yml at the beginning of docs easily screws up the entire html rendering

Control if this file starts numbering

3.1 numbering: start-at: 1 # Start at Section 1 level: 1 # Chapter level

Introduction

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

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

`[x] introduces and motivates the core question or problem`

4.1 The Coordination Crisis in AI Governance

As AI capabilities advance at an accelerating pace—demonstrated by the rapid progression from GPT-3 to GPT-4, Claude, and beyond—we face a governance challenge unlike any in human history: how to ensure increasingly powerful AI systems remain aligned with human values and beneficial to humanity's long-term flourishing. This challenge becomes particularly acute when considering the possibility of transformative AI systems that could drastically alter civilization's trajectory, potentially including existential risks from misaligned systems.

> Despite unprecedented investment in AI safety research, rapidly growing awareness among key stakeholders, and proliferating frameworks for responsible AI development, we face what I'll term the "coordination crisis" in AI governance—a systemic failure to align diverse efforts across technical, policy, and strategic domains into a coherent response proportionate to the risks we face.

The AI governance landscape exhibits a peculiar paradox: extraordinary activity alongside fundamental coordination failure. Consider the current state of affairs:

Technical safety researchers develop increasingly sophisticated alignment techniques, but often

without clear implementation pathways to deployment contexts. Policy specialists craft principles and regulatory frameworks without sufficient technical grounding to ensure their practical efficacy. Ethicists articulate normative principles that lack operational specificity. Strategy researchers identify critical uncertainties but struggle to translate these into actionable guidance.

4.1.1 Empirical Paradox: Investment Alongside Fragmentation

The fragmentation problem manifests in incompatible frameworks between technical researchers, policy specialists, and strategic analysts. Each community develops sophisticated approaches within their domain, yet translation between domains remains primitive. This creates systematic blind spots where risks emerge at the interfaces between technical capabilities, institutional responses, and strategic dynamics.

4.1.2 Systematic Risk Increase Through Coordination Failure

Coordination failures systematically amplify existential risk through multiple pathways. Safety gaps emerge when technical solutions lack policy implementation pathways. Resource misallocation occurs when multiple teams unknowingly duplicate efforts while critical areas remain unaddressed. Most perniciously, locally optimized decisions by individual actors can create negative-sum dynamics that increase overall risk—a AI governance tragedy of the commons.

4.1.3 Historical Parallels and Temporal Urgency

Traditional governance approaches evolved for technologies with longer development cycles and clearer deployment boundaries. The nuclear era provided decades for international regime development. Climate governance, despite its challenges, addresses a phenomenon unfolding over centuries. AI development, by contrast, may transition from current capabilities to transformative systems within years or decades, compressing the available window for effective coordination.

4.2 Research Question and Scope

This thesis addresses a specific dimension of the coordination challenge by investigating the question: **Can frontier AI technologies be utilized to automate the modeling of transformative AI risks, enabling robust prediction of policy impacts across diverse worldviews?**

Refined Thesis Statement: This thesis demonstrates that frontier language models can automate the extraction and formalization of probabilistic world models from AI safety literature, creating a scalable computational framework that enhances coordination in AI governance through systematic policy evaluation under uncertainty.

To break this down into its components:

> **Frontier AI Technologies**: Today's most capable language models (GPT-4, Claude-3 level systems)

> **Automated Modeling**: Using these systems to extract and formalize argument structures from natural language
> **Transformative AI Risks**: Potentially catastrophic outcomes from advanced AI systems, particularly existential risks
> **Policy Impact Prediction**: Evaluating how governance interventions might alter probability distributions over outcomes
> **Diverse Worldviews**: Accounting for fundamental disagreements about AI development trajectories and risk factors

The investigation encompasses both theoretical development and practical implementation, focusing specifically on existential risks from misaligned AI systems rather than broader AI ethics concerns. This narrowed scope enables deep technical development while addressing the highest-stakes coordination challenges.

4.3 The Multiplicative Benefits Framework

The central thesis of this work is that combining three elements—automated worldview extraction, prediction market integration, and formal policy evaluation—creates multiplicative rather than merely additive benefits for AI governance. Each component enhances the others, creating a system more valuable than the sum of its parts.

Automated worldview extraction using frontier language models addresses the scaling bottleneck in current approaches to AI risk modeling. The Modeling Transformative AI Risks (MTAIR) project demonstrated the value of formal representation but required extensive manual effort to translate qualitative arguments into quantitative models. Automation enables processing orders of magnitude more content, incorporating diverse perspectives, and maintaining models in near real-time as new arguments emerge.

Prediction market integration grounds these models in collective forecasting intelligence. By connecting formal representations to live forecasting platforms, the system can incorporate timely judgments about critical uncertainties from calibrated forecasters. This creates a dynamic feedback loop, where models inform forecasters and forecasts update models.

Formal policy evaluation transforms static risk assessments into actionable guidance by modeling how specific interventions might alter critical parameters. This enables conditional forecasting—understanding not just the probability of adverse outcomes but how those probabilities change under different policy regimes.

The synergy emerges because automation enables comprehensive data integration, markets inform and validate models, and evaluation gains precision from both automated extraction and market-based calibration.

4.4 Thesis Structure and Roadmap

The remainder of this thesis develops the multiplicative benefits framework from theoretical foundations to practical implementation, following a progression from abstract principles to

concrete applications:

Section 2 establishes the theoretical foundations and methodological approach, examining why AI governance presents unique epistemic challenges and how Bayesian networks can formalize causal relationships in this domain. This section grounds the technical contributions in established theory while identifying the specific gaps AMTAIR addresses.

Section 3 presents the AMTAIR implementation, detailing the technical system that transforms qualitative arguments into formal representations. It demonstrates the approach through two case studies: the canonical Rain-Sprinkler-Lawn example for intuitive understanding and the more complex Carlsmith model of power-seeking AI for real-world validation.

Section 4 provides critical analysis of the approach, addressing potential failure modes, scaling challenges, and integration with existing governance frameworks. This section engages seriously with objections and limitations while demonstrating the robustness of the core approach.

Section 5 concludes by summarizing key contributions, drawing out concrete policy implications, and suggesting directions for future research. It returns to the opening coordination crisis to show how AMTAIR provides partial but significant solutions.

Throughout this progression, I maintain a dual focus on theoretical sophistication and practical utility. The framework aims not merely to advance academic understanding of AI risk but to provide actionable tools for improving coordination in AI governance.

Having established the coordination crisis and outlined how automated modeling can address it, we now turn to the theoretical foundations that make this approach possible. The next chapter examines the unique epistemic challenges of AI governance and introduces the formal tools—particularly Bayesian networks—that enable rigorous reasoning under deep uncertainty.

Context & Background

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

> 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

5.1 Theoretical Foundations

5.1.1 AI Existential Risk: The Carlsmith Model

Carlsmith's "Is power-seeking AI an existential risk?" (2021) represents one of the most structured approaches to assessing the probability of existential catastrophe from advanced AI. The analysis decomposes the overall risk into six key premises, each with an explicit probability estimate.

> Carlsmith [2] provides the canonical structured approach to AI existential risk assessment

Six-Premise Decomposition:

Carlsmith decomposes existential risk into a probabilistic chain with explicit estimates:

1. **Premise 1**: Transformative AI development this century (P 0.80)
2. **Premise 2**: AI systems pursuing objectives in the world (P 0.95)
3. **Premise 3**: Systems with power-seeking instrumental incentives (P 0.40)
4. **Premise 4**: Sufficient capability for existential threat (P 0.65)
5. **Premise 5**: Misaligned systems despite safety efforts (P 0.50)
6. **Premise 6**: Catastrophic outcomes from misaligned power-seeking (P 0.65)

Composite Risk Calculation: P(doom) 0.05 (5%)

This structured approach exemplifies the type of reasoning that AMTAIR aims to formalize and automate, providing both transparency in assumptions and modularity for critique and refinement.

5.1.1.1 Why Carlsmith as Ideal Formalization Target

Carlsmith's model represents "low-hanging fruit" for automated formalization because it already exhibits explicit probabilistic reasoning with clear conditional dependencies. Success with this structured argument validates the approach for less explicit arguments throughout AI safety literature. The model demonstrates several key features that make it ideal for formalization: explicitly probabilistic reasoning with quantified estimates, clear conditional dependencies between premises, transparent decomposition of complex causal pathways, well-documented argumentation available for extraction validation, and policy-relevant implications requiring formal evaluation.

5.1.2 The Epistemic Challenge of Policy Evaluation

AI governance policy evaluation faces unique epistemic challenges that render traditional policy analysis methods insufficient. The domain combines complex causal chains with limited empirical grounding, deep uncertainty about future capabilities, divergent stakeholder worldviews, and few opportunities for experimental testing before deployment.

Traditional methods fall short in several ways. Cost-benefit analysis struggles with existential outcomes and deep uncertainty about unprecedented events. Scenario planning often lacks the probabilistic reasoning necessary for rigorous evaluation under uncertainty. Expert elicitation alone fails to formalize interdependencies between variables and make assumptions explicit. Qualitative approaches obscure crucial assumptions that drive conclusions, making it difficult to identify cruxes of disagreement.

Unprecedented Epistemic Environment:

The AI governance domain presents specific challenges that traditional policy analysis cannot adequately address:

> **Deep Uncertainty**: Many decisions involve unprecedented scenarios without historical frequency data for calibration
> **Complex Causality**: Policy effects propagate through multi-level dependencies spanning technical, institutional, and strategic domains
> **Multidisciplinary Integration**: Combining technical facts, ethical principles, and strategic considerations requires novel synthesis approaches
> **Value-Laden Assessment**: Risk evaluation inherently involves normative judgments about acceptable outcomes and distributional effects

5.1.2.1 Unique Difficulties in AI Governance

Complex Causal Chains: Multi-level dependencies between technical capabilities, institutional responses, and strategic outcomes create analytical challenges beyond traditional policy domains.

Deep Uncertainty: Unprecedented AI capabilities make historical analogies insufficient, requiring new approaches to reasoning about low-probability, high-impact events.

Divergent Worldviews: Fundamental disagreements persist about timeline expectations for transformative AI, difficulty of alignment problems, effectiveness of governance interventions, and possibilities for international coordination.

5.1.2.2 Limitations of Traditional Policy Analysis

Traditional policy analysis approaches prove inadequate for AI governance challenges. Cost-benefit analysis struggles with potentially infinite expected values from existential outcomes and lacks frameworks for deep uncertainty. Scenario planning, while useful for exploration, often lacks the probabilistic reasoning necessary for rigorous uncertainty quantification and policy comparison. Expert elicitation methods fail to formalize complex interdependencies between variables, leaving implicit assumptions unexamined. Qualitative frameworks, though rich in insight, obscure crucial assumptions and parameter sensitivities that drive different conclusions about optimal policies.

5.1.3 Argument Mapping and Formal Representations

Argument mapping offers a bridge between informal reasoning in natural language and the formal representations needed for rigorous analysis. By explicitly identifying claims, premises, inferential relationships, and support/attack patterns, argument maps make implicit reasoning structures visible for examination and critique.

The progression from natural language arguments to formal Bayesian networks requires an intermediate representation that preserves narrative structure while adding mathematical precision. The ArgDown format serves this purpose by encoding hierarchical relationships between statements, while its extension, BayesDown, adds probabilistic metadata to enable full Bayesian network construction.

```
[Effect_Node]: Description of effect. {"instantiations": ["effect_TRUE", "effect_FALSE"]}
 + [Cause_Node]: Description of direct cause. {"instantiations": ["cause_TRUE", "cause_FALSE
   + [Root_Cause]: Description of indirect cause. {"instantiations": ["root_TRUE", "root_FAL
```

5.1.4 Bayesian Networks as Knowledge Representation

Bayesian networks provide a formal mathematical framework for representing causal relationships and reasoning under uncertainty. These directed acyclic graphs (DAGs) combine qualitative structure—nodes representing variables and edges representing dependencies—with quantitative parameters in the form of conditional probability tables.

5.1.4.1 Mathematical Foundations

Bayesian networks provide a formal mathematical framework for representing causal relationships and reasoning under uncertainty through Directed Acyclic Graphs (DAGs) combining qualitative structure with quantitative parameters.

Core Components:

> **Nodes**: Variables with discrete states representing propositions or factors
> **Edges**: Directed relationships representing conditional dependencies
> **Acyclicity**: Ensuring coherent probabilistic interpretation without circular dependencies
> **Conditional Probability Tables**: Quantifying P(Node|Parents) for all parent state combinations

Probability Factorization: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$

5.1.4.2 The Rain-Sprinkler-Grass Example

This simple example demonstrates all key concepts while remaining intuitive. The network structure consists of Rain as a root cause with P(rain) = 0.2, Sprinkler as an intermediate variable where P(sprinkler|rain) varies by rain state, and Grass_Wet as the effect where P(wet|rain, sprinkler) depends on both causes.

The example enables various inference capabilities including marginal probabilities such as P(grass_wet) computed from the joint distribution, conditional queries like P(rain|grass_wet) for diagnostic reasoning, and counterfactual analysis such as P(grass_wet|do(sprinkler=false)) for intervention effects.

```
# Basic network representation
nodes = ['Rain', 'Sprinkler', 'Grass_Wet']
edges = [('Rain', 'Sprinkler'), ('Rain', 'Grass_Wet'), ('Sprinkler', 'Grass_Wet')]

# Conditional probability specification
P_wet_given_causes = {
    (True, True): 0.99,    # Rain=T, Sprinkler=T
    (True, False): 0.80,   # Rain=T, Sprinkler=F
    (False, True): 0.90,   # Rain=F, Sprinkler=T
    (False, False): 0.01   # Rain=F, Sprinkler=F
}
```

5.1.4.3 Advantages for AI Risk Modeling

Bayesian networks offer several key advantages for AI risk modeling. They provide explicit uncertainty representation where all beliefs are represented with probability distributions rather than point estimates. The framework naturally supports causal reasoning through native support for intervention analysis and counterfactual reasoning via do-calculus. Evidence integration becomes principled through Bayesian updating mechanisms. The modular structure allows com-

plex arguments to be decomposed into manageable, verifiable components. Finally, the visual communication provided by graphical representation facilitates understanding across different expertise levels.

5.1.5 The MTAIR Framework: Achievements and Limitations

The Modeling Transformative AI Risks (MTAIR) project demonstrated the value of formal probabilistic modeling for AI safety, but also revealed significant limitations in the manual approach. While MTAIR successfully translated complex arguments into Bayesian networks and enabled sensitivity analysis, the intensive human labor required for model creation limited both scalability and timeliness.

> Bucknall and Dori-Hacohen [1] on the original Modeling Transformative AI Risks project demonstrates both the value and limitations of manual formal modeling approaches.

5.1.5.1 MTAIR's Innovations

MTAIR's key innovations advanced the field of AI risk modeling significantly. The project introduced structured uncertainty representation through explicit probability distributions over key variables rather than point estimates. It developed systematic methods for expert judgment integration, aggregating diverse expert opinions and beliefs. The sensitivity analysis capabilities enabled identification of critical uncertainties that most significantly drive overall conclusions. Perhaps most importantly, it established direct connections between technical risk models and governance implications, bridging the gap between technical analysis and policy application.

5.1.5.2 Fundamental Limitations Motivating AMTAIR

Despite its innovations, MTAIR faces fundamental limitations that motivate the automated approach. The scalability bottleneck is severe—manual model construction requires weeks of expert effort per argument, making comprehensive coverage impossible. The static nature of manually constructed models provides no mechanisms for updating as new research and evidence emerge. Limited accessibility restricts usage to specialists with formal modeling expertise, excluding many stakeholders. Finally, the single worldview focus creates difficulty in representing multiple conflicting perspectives simultaneously, limiting the framework's utility for coordination across diverse viewpoints.

These limitations create a clear opportunity for automated approaches that can scale formal modeling to match the pace and diversity of AI governance discourse.

5.1.5.3 Mechanics of World Modeling in Analytica

The MTAIR project's Analytica implementation provides important lessons for automation. The manual process involves several key steps: variable identification through careful reading of source texts, structure elicitation via expert interviews and workshops, probability quantification using various elicitation techniques, and validation through sensitivity analysis and expert review.

Each step requires significant time and expertise, with a single model taking weeks to months to develop. Understanding these mechanics helps identify specific opportunities for automation while preserving the rigor of the manual approach.

5.1.6 Literature Review: Content Level

5.1.6.1 AI Risk Models Evolution

The evolution of AI risk models reflects increasing sophistication in both structure and quantification. Early models focused on simple binary outcomes, while recent work incorporates complex causal chains and continuous variables. Key developments include:

The progression from qualitative arguments to structured probabilistic models demonstrates the field's maturation and the increasing recognition that rigorous quantitative analysis is essential for policy evaluation.

5.1.6.2 Governance Proposals Taxonomy

AI governance proposals can be categorized along several dimensions:

> **Technical Standards**: Safety requirements, testing protocols, capability thresholds
> **Regulatory Frameworks**: Licensing regimes, liability structures, oversight mechanisms
> **International Coordination**: Treaties, soft law arrangements, technical cooperation
> **Research Priorities**: Funding allocation, talent development, knowledge sharing

5.1.7 Literature Review: Technical/Theoretical Background

5.1.7.1 Bayesian Network Theory

The theoretical foundations of Bayesian networks rest on probability theory and graph theory. Key concepts include conditional independence encoded through d-separation, the Markov condition relating graph structure to probabilistic relationships, and inference algorithms ranging from exact methods like variable elimination to approximate approaches like Monte Carlo sampling.

5.1.7.2 Software Tools Landscape

The implementation of AMTAIR builds on established software libraries:

> **pgmpy**: Python library for probabilistic graphical models, providing network construction and inference
> **NetworkX**: Graph analysis and manipulation capabilities
> **PyVis**: Interactive network visualization
> **Pandas/NumPy**: Data manipulation and numerical computation

5.1.7.3 Formalization Approaches

Formalizing natural language arguments into mathematical models involves several theoretical challenges. The translation must preserve semantic content while adding mathematical preci-

sion. Key approaches include structured extraction templates, semantic parsing techniques, and hybrid human-AI workflows.

5.1.7.4 Correlation Accounting Methods

Standard Bayesian networks assume conditional independence given parents, but real-world AI risk factors often exhibit complex correlations. Methods for handling correlations include:

> **Copula Methods**: Modeling dependence structures separately from marginal distributions
> **Hierarchical Models**: Capturing correlations through shared latent variables
> **Explicit Correlation Nodes**: Adding nodes to represent correlation mechanisms
> **Sensitivity Bounds**: Analyzing impact of independence assumptions

5.2 Methodology

5.2.1 Research Design Overview

This research combines theoretical development with practical implementation, following an iterative approach that moves between conceptual refinement and technical validation. The methodology encompasses formal framework development, computational implementation, extraction quality assessment, and application to real-world AI governance questions.

The research process follows four integrated phases:

1. **Framework Development**: Creating theoretical foundations for automated worldview extraction
2. **Technical Implementation**: Building computational tools as working prototype
3. **Empirical Validation**: Assessing quality against expert benchmarks
4. **Policy Application**: Demonstrating practical utility for governance questions

5.2.2 Formalizing World Models from AI Safety Literature

The core methodological challenge involves transforming natural language arguments in AI safety literature into formal causal models with explicit probability judgments. This extraction process identifies key variables, causal relationships, and both explicit and implicit probability estimates through a systematic pipeline.

The extraction approach combines several elements: identification of key variables and entities in text, recognition of causal claims and relationships, detection of explicit and implicit probability judgments, transformation into structured intermediate representations, and conversion to formal Bayesian networks.

Large language models facilitate this process through specialized techniques including two-stage prompting that separates structure from probability extraction, specialized templates for different types of source documents, techniques for identifying implicit assumptions and relationships, and mechanisms for handling ambiguity and uncertainty.

5.2.3 From Natural Language to Computational Models

5.2.3.1 The Two-Stage Extraction Process

AMTAIR employs a novel two-stage process that separates structural argument extraction from probability quantification, enabling modular improvement and human oversight at critical decision points.

Stage 1: Structural Extraction (ArgDown Generation)

The first stage focuses on identifying the argument structure: extracting key propositions and entities from natural language text, mapping support/attack relationships and conditional dependencies, constructing properly nested argument representations that preserve logical flow, and creating ArgDown format suitable for both human review and machine processing.

```
def extract_argument_structure(text):
    """Extract hierarchical argument structure from natural language"""
    # LLM-based extraction with specialized prompts
    prompt = ArgumentExtractionPrompt(
        text=text,
        output_format="ArgDown",
        focus_areas=["causal_claims", "probability_statements", "conditional_reasoning"]
    )

    structure = llm.complete(prompt)
    return validate_argdown_syntax(structure)
```

Stage 2: Probability Integration (BayesDown Enhancement)

The second stage adds quantitative information: identifying and parsing numerical probability statements in source text, creating systematic elicitation questions for implicit probability judgments, incorporating domain expertise for ambiguous or missing quantifications, and ensuring probability assignments satisfy basic coherence requirements.

```
def integrate_probabilities(argdown_structure, probability_sources):
    """Convert ArgDown to BayesDown with probabilistic information"""
    questions = generate_probability_questions(argdown_structure)
    probabilities = extract_probabilities(probability_sources, questions)

    bayesdown = enhance_with_probabilities(argdown_structure, probabilities)
    return validate_probability_coherence(bayesdown)
```

5.2.4 Directed Acyclic Graphs: Structure and Semantics

Directed Acyclic Graphs (DAGs) form the mathematical foundation of Bayesian networks, encoding both the qualitative structure of causal relationships and the quantitative parameters that define conditional dependencies. In AI risk modeling, these structures represent causal

pathways to potential outcomes of interest.

Key mathematical properties essential for AI risk modeling include the acyclicity requirement ensuring coherent probabilistic interpretation without logical contradictions, d-separation defining conditional independence relationships between variables based on graph structure, the Markov condition where each variable is conditionally independent of non-descendants given parents, and path analysis revealing causal pathways and information flow through the network structure.

The causal interpretation in AI governance contexts follows Pearl's framework, where edges represent direct causal influence between factors, intervention analysis through do-calculus enables rigorous evaluation of policy effects, counterfactual reasoning supports "what if" scenarios essential for governance planning, and evidence integration through Bayesian updating incorporates new information and expert judgment.

5.2.5 Quantification of Probabilistic Judgments

Transforming qualitative uncertainty expressions into quantitative probabilities requires systematic interpretation frameworks that account for individual and cultural variation.

Standard linguistic mappings (with significant individual variation) include:

> "Very likely" → 0.8-0.9
> "Probable" → 0.6-0.8
> "Uncertain" → 0.4-0.6
> "Unlikely" → 0.2-0.4
> "Highly improbable" → 0.05-0.15

Expert elicitation methodologies provide various approaches: direct probability assessment asking "What is P(outcome)?" with calibration training, comparative assessment asking "Is A more likely than B?" for relative judgment validation, frequency format asking "In 100 similar cases, how many would result in outcome?" for clearer mental models, and betting odds asking "What odds would you accept for this bet?" for revealed preference elicitation.

Calibration and validation face several challenges including individual variation in linguistic interpretation and probability anchoring, domain-specific anchoring and reference class selection, cultural and contextual influences on uncertainty expression and tolerance, and limited empirical basis for calibration in unprecedented scenarios like transformative AI.

5.2.6 Inference Techniques for Complex Networks

Once Bayesian networks are constructed, probabilistic inference enables reasoning about uncertainties, counterfactuals, and policy interventions. For the complex networks representing AI risks, computational approaches must balance accuracy with tractability.

Inference methods implemented include exact methods for smaller networks (variable elimination, junction trees), approximate methods for larger networks (Monte Carlo sampling, variational inference), specialized approaches for rare event analysis, and intervention modeling for policy evaluation using do-calculus.

Implementation considerations involve computational complexity management through network decomposition, sampling efficiency optimization via importance sampling, approximation quality monitoring with convergence diagnostics, and uncertainty representation in outputs including confidence intervals.

5.2.7 Integration with Prediction Markets and Forecasting Platforms

To maintain relevance in a rapidly evolving field, formal models must integrate with live data sources such as prediction markets and forecasting platforms. This integration enables continuous updating of model parameters as new information emerges.

Live data sources for dynamic model updating include:

> **Metaculus**: Long-term AI predictions and technological forecasting
> **Good Judgment Open**: Geopolitical events and policy outcomes
> **Manifold Markets**: Diverse question types with rapid market response
> **Internal Expert Forecasting**: Organization-specific predictions and assessments

The data processing and integration pipeline connects these sources:

```
def integrate_forecast_data(model_variables, forecast_platforms):
    """Connect Bayesian network variables to live forecasting data"""
    mappings = create_semantic_mappings(model_variables, forecast_platforms)

    for variable, forecasts in mappings.items():
        weighted_forecast = aggregate_forecasts(
            forecasts,
            weights=calculate_track_record_weights(forecasts)
        )
        model.update_prior(variable, weighted_forecast)

    return model.recompute_posteriors()
```

Technical implementation challenges include question mapping to connect forecast questions to specific model variables with semantic accuracy, temporal alignment handling different forecast horizons and update frequencies, conflict resolution through principled aggregation when sources provide contradictory information, and track record weighting incorporating forecaster calibration and expertise into aggregation.

With these theoretical foundations and methodological approaches established, we can now present the AMTAIR system implementation. The next chapter demonstrates how these concepts translate into a working prototype that automates the extraction and formalization of world models from AI safety literature.

AMTAIR Implementation

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

> 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

6.1 Software Implementation

6.1.1 System Architecture and Data Flow

The AMTAIR system implements an end-to-end pipeline from unstructured text to interactive Bayesian network visualization. Its modular architecture comprises five main components that progressively transform information from natural language into formal models suitable for policy analysis.

The five-stage pipeline architecture demonstrates how each component builds on the previous, with validation checkpoints preventing error propagation:

1. **Text Ingestion and Preprocessing**: Handles format normalization (PDF, HTML, Markdown), metadata extraction, citation tracking, and relevance filtering
2. **BayesDown Extraction**: Two-stage argument structure identification and probabilistic information integration with quality validation
3. **Structured Data Transformation**: Parsing into standardized relational formats with network topology validation
4. **Bayesian Network Construction**: Mathematical model instantiation using NetworkX and pgmpy libraries
5. **Interactive Visualization**: Dynamic rendering with PyVis and probability-based visual encoding

```python
class AMTAIRPipeline:
    def __init__(self):
        self.ingestion = DocumentIngestion()
        self.extraction = BayesDownExtractor()
        self.transformation = DataTransformer()
        self.network_builder = BayesianNetworkBuilder()
        self.visualizer = InteractiveVisualizer()

    def process(self, document):
        """End-to-end processing from document to interactive model"""
        structured_data = self.ingestion.preprocess(document)
        bayesdown = self.extraction.extract(structured_data)
        dataframe = self.transformation.convert(bayesdown)
        network = self.network_builder.construct(dataframe)
        return self.visualizer.render(network)
```

The design principles emphasize scalability through modular architecture where each component can be improved independently, standard interfaces using JSON and CSV formats for interoperability, validation checkpoints with quality gates at each stage, and an extensible framework supporting additional analysis capabilities without core changes.

6.1.2 Rain-Sprinkler-Grass Example Implementation

The Rain-Sprinkler-Grass example serves as a canonical test case demonstrating each step in the AMTAIR pipeline. This simple causal scenario—where both rain and sprinkler use can cause wet grass, and rain influences sprinkler use—provides an intuitive introduction to Bayesian network concepts while exercising all system components.

Stage 1: BayesDown Input Representation

The structured representation captures both hierarchical relationships and probability information:

```
[Grass_Wet]: Concentrated moisture on, between and around the blades of grass.
{"instantiations": ["grass_wet_TRUE", "grass_wet_FALSE"],
 "priors": {"p(grass_wet_TRUE)": "0.322", "p(grass_wet_FALSE)": "0.678"},
 "posteriors": {
   "p(grass_wet_TRUE|sprinkler_TRUE,rain_TRUE)": "0.99",
   "p(grass_wet_TRUE|sprinkler_TRUE,rain_FALSE)": "0.9",
   "p(grass_wet_TRUE|sprinkler_FALSE,rain_TRUE)": "0.8",
   "p(grass_wet_TRUE|sprinkler_FALSE,rain_FALSE)": "0.0"
 }}
 + [Rain]: Tears of angels crying high up in the skies hitting the ground.
   {"instantiations": ["rain_TRUE", "rain_FALSE"],
    "priors": {"p(rain_TRUE)": "0.2", "p(rain_FALSE)": "0.8"}}
```

```
+ [Sprinkler]: Activation of a centrifugal force based CO2 droplet distribution system.
  {"instantiations": ["sprinkler_TRUE", "sprinkler_FALSE"],
   "priors": {"p(sprinkler_TRUE)": "0.44838", "p(sprinkler_FALSE)": "0.55162"},
   "posteriors": {
     "p(sprinkler_TRUE|rain_TRUE)": "0.01",
     "p(sprinkler_TRUE|rain_FALSE)": "0.4"
   }}
  + [Rain]
```

Stage 2: Automated Parsing and Data Extraction

The parsing algorithm (`parse_markdown_hierarchy_fixed`) processes the BayesDown format to extract structured information. The algorithm removes comments and cleans text, extracts titles, descriptions, and indentation levels, establishes parent-child relationships based on indentation following BayesDown semantics, converts to DataFrame format with all necessary columns, and adds derived columns for network analysis such as node types and Markov blankets.

Stage 3: Bayesian Network Construction and Validation

Network construction transforms the DataFrame into a formal Bayesian network by creating directed graph structure using NetworkX, adding nodes with complete probabilistic information, establishing edges based on extracted parent-child relationships, validating DAG properties to ensure acyclicity, and preparing for inference with conditional probability tables.

Stage 4: Interactive Visualization with Probability Encoding

The visualization strategy employs multiple visual channels to convey information: node colors using a green (high probability) to red (low probability) gradient based on primary state likelihood, border colors with blue for root nodes, purple for intermediate nodes, and magenta for leaf nodes, clear edge directions showing causal influence, and interactive elements including click actions for detailed probability tables and drag functionality for layout adjustment.

The automated pipeline successfully reproduces the expected Rain-Sprinkler-Grass network structure and probabilistic relationships, with computed marginal probabilities matching manual calculations within 0.001 precision, validating the extraction and transformation processes.

6.1.3 Carlsmith Implementation

Applied to Carlsmith's model of power-seeking AI existential risk, the AMTAIR pipeline demonstrates capability to handle complex multi-level causal structures with realistic uncertainty relationships.

Model Complexity and Scope:

The Carlsmith model represents a significant increase in complexity:

- **23 nodes** representing AI development factors and risk pathways
- **45 conditional dependencies** capturing complex causal relationships

> **6 primary risk pathways** to existential catastrophe outcomes
> **Multiple temporal stages** from capability development through deployment to outcome

Core Risk Pathway Structure:

```
Existential_Catastrophe ← Human_Disempowerment ← Scale_Of_Power_Seeking
                                                ← Misaligned_Power_Seeking
                                                ← [APS_Systems, Difficulty_Of_Alignment, Dep
```

Advanced BayesDown Representation Example:

```
{
  "instantiations": ["misaligned_power_seeking_TRUE", "misaligned_power_seeking_FALSE"],
  "priors": {"p(misaligned_power_seeking_TRUE)": "0.338"},
  "posteriors": {
    "p(misaligned_power_seeking_TRUE|aps_systems_TRUE, difficulty_of_alignment_TRUE, deploym
    "p(misaligned_power_seeking_TRUE|aps_systems_TRUE, difficulty_of_alignment_FALSE, deploy
    "p(misaligned_power_seeking_TRUE|aps_systems_FALSE, difficulty_of_alignment_TRUE, deploy
  }
}
```

Sensitivity Analysis Results:

The implementation enables identification of critical variables with highest impact on final outcome:

1. **APS_Systems development** (probability range affects outcome by 40%)
2. **Difficulty_Of_Alignment assessment** (30% outcome variation)
3. **Deployment_Decisions under uncertainty** (25% outcome variation)

Intervention Analysis demonstrates policy evaluation capabilities:

> Preventing APS deployment reduces P(catastrophe) from 5% to 0.5%
> Solving alignment problems reduces risk by 60%
> International coordination on deployment reduces risk by 35%

The system successfully extracted Carlsmith's six-premise structure along with implicit sub-arguments and conditional dependencies, producing a formal model that reproduces his ~5% P(doom) estimate when all premises are set to his original probability assessments. Implementation performance metrics show extraction time of ~3 minutes for complete document processing, network construction in <10 seconds for the 23-node network, millisecond response time for standard probabilistic queries, and 94% agreement with manual expert annotation of argument structure.

6.1.4 Inference & Extensions

Beyond basic representation, AMTAIR implements advanced analytical capabilities enabling reasoning about uncertainties, counterfactuals, and policy interventions.

6.1.4.1 Probabilistic Inference Engine

The system supports multiple query types essential for policy analysis:

```
# Marginal probability queries for outcomes of interest
P_catastrophe = network.query
```

(['Existential_Catastrophe'])

Conditional probability queries given evidence

P_catastrophe_given_aps = network.query([‘Existential_Catastrophe’], evidence={‘APS_Systems’: ‘aps_systems_TRUE’})

Intervention analysis using do-calculus for policy evaluation

P_catastrophe_no_deployment = network.do_query('Deployment_Decisions', 'WITHHOLD', ['Existential_Catastrophe'])

```
Algorithm selection adapts to network complexity: exact methods using variable elimination f

#### Policy Evaluation Interface {#sec-policy-evaluation}

The policy evaluation framework enables systematic assessment of governance interventions:

```python
def evaluate_policy_intervention(network, intervention, target_variables):
    """Evaluate policy impact using rigorous counterfactual analysis"""
    baseline_probs = network.query(target_variables)
    intervention_probs = network.do_query(intervention['variable'],
                                          intervention['value'],
                                          target_variables)

    return {
        'baseline': baseline_probs,
        'intervention': intervention_probs,
        'effect_size': compute_effect_size(baseline_probs, intervention_probs),
        'robustness': assess_robustness_across_scenarios(intervention)
    }
```

Example policy evaluations demonstrate practical applications including compute governance through restricting access to large-scale computing resources, safety standards via mandatory testing before deployment, and international coordination through binding agreements on development pace.
```

8.0.0.1 Extensions and Future Capabilities

Prediction Market Integration (partially implemented): - Real-time probability updates from Metaculus and other platforms - Question mapping between forecasts and model variables - Automated relevance scoring and confidence weighting

Cross-Worldview Analysis capabilities: - Multiple model comparison and consensus identification - Crux analysis highlighting key disagreements - Robust strategy identification across uncertainty

Sensitivity Analysis Implementation provides critical insights: - Identification of parameters driving outcome uncertainty - Visualization of parameter influence on conclusions - Guidance for targeted research priorities

8.1 Results

8.1.1 Extraction Quality Assessment

Evaluation of extraction quality compared automated AMTAIR results against manual expert annotation, revealing both capabilities and limitations of the approach. Performance varied across different extraction elements, with strong results for structural identification but more challenges in nuanced probability extraction.

Preliminary Performance Indicators (based on initial testing):

Structural extraction shows promising results with node identification achieving high accuracy for clearly defined entities, relationship extraction performing well for explicit causal language, and hierarchy construction correctly capturing most parent-child relationships.

Probability extraction faces greater challenges, with explicit probability statements extracted accurately when numerical values are clearly stated, qualitative expressions showing more variation in interpretation, and complex conditional relationships requiring iterative refinement.

Error Analysis and Pattern Recognition:

Common extraction challenges include: - **Implicit Assumptions**: Unstated background assumptions requiring domain knowledge - **Complex Conditionals**: Nested "if-then" statements with multiple interacting conditions - **Ambiguous Quantifiers**: Terms like "significant" or "likely" without clear context - **Cross-Reference Resolution**: Pronouns and indirect references requiring disambiguation

Successful extraction occurs most reliably with clear causal language ("X causes Y", "leads to"), explicit probability statements containing numerical values, simple conditional structures with clear antecedents, and well-structured arguments using standard premise indicators.

8.1.2 Computational Performance Analysis

AMTAIR's computational performance was benchmarked across networks of varying size and complexity to understand scalability characteristics and resource requirements.

Scaling Performance Characteristics:

Network size significantly impacts processing time: - Small networks (10 nodes): <1 second end-to-end processing - Medium networks (11-30 nodes): 2-8 seconds total processing time - Large networks (31-50 nodes): 15-45 seconds total processing time - Very large networks (>50 nodes): Require approximate inference methods

Component-Level Performance Analysis:

Each pipeline stage exhibits different scaling characteristics. BayesDown parsing shows O(n) linear scaling with document length, remaining efficient even for long documents. Network construction exhibits $O(n^2)$ scaling with number of variables and relationships, becoming the primary bottleneck for large networks. Visualization rendering scales as O(n + e) with nodes and edges, requiring optimization for networks exceeding 50 nodes. Exact inference faces exponential worst-case complexity but demonstrates polynomial typical-case performance for sparse networks common in AI risk models.

Memory and resource requirements vary by model complexity, with peak memory usage ranging from 2-8 GB for complex models during network construction, storage requirements of 10-50 MB per complete model including visualizations, and API costs of $0.10-0.50 per document for LLM-based extraction using GPT-4 class models.

8.1.3 Case Study: The Carlsmith Model Formalized

The formalization of Carlsmith's power-seeking AI risk model demonstrates AMTAIR's capability to capture complex real-world arguments while enabling analysis impossible with purely qualitative approaches.

Formalized Model Characteristics:

The extracted model successfully represents: - **21 distinct variables** capturing main premises and detailed sub-components - **27 directional relationships** representing causal connections and dependencies - **Complete CPT specification** for all conditional probability relationships - **Preserved semantic content** from original argument while enabling formal analysis - **Validated aggregate calculation** reproducing Carlsmith's ~5% existential risk estimate

Structural Insights from Formalization:

Network analysis reveals important properties of the argument structure:

```
network_metrics = {
    'nodes': 21,
    'edges': 27,
    'max_path_length': 6,  # Longest causal chain from root to outcome
    'branching_factor': 2.3,  # Average number of children per parent
    'root_nodes': 8,  # Variables with no parents (exogenous factors)
    'leaf_nodes': 1   # Variables with no children (final outcome)
}
```

Sensitivity Analysis Results:

Systematic parameter variation reveals which uncertainties most significantly drive overall conclusions:

1. **APS_Systems Development** (±0.4 probability range affects outcome by 40%)
2. **Difficulty_Of_Alignment Assessment** (30% outcome variation range)
3. **Deployment_Decisions Under Uncertainty** (25% outcome variation range)
4. **Corrective_Feedback Effectiveness** (20% outcome variation range)

Policy Intervention Analysis:

The formalized model enables rigorous evaluation of potential interventions:

```
intervention_results = {
    'prevent_aps_deployment': {
        'baseline_risk': 0.05,
        'intervention_risk': 0.005,
        'relative_reduction': 0.90
    },
    'solve_alignment_problems': {
        'baseline_risk': 0.05,
        'intervention_risk': 0.02,
        'relative_reduction': 0.60
    },
    'international_coordination': {
        'baseline_risk': 0.05,
        'intervention_risk': 0.035,
        'relative_reduction': 0.30
    }
}
```

8.1.4 Comparative Analysis of AI Governance Worldviews

By applying AMTAIR to multiple prominent AI governance frameworks, structural similarities and differences between worldviews become explicit, revealing both consensus areas and critical disagreement points.

Multi-Perspective Extraction Results:

Analysis of three representative worldviews reveals systematic differences:

Variable	Technical Optimists	Governance Advocates	Alignment Researchers	Std Dev
AI Timeline	15-30 years	10-20 years	5-15 years	0.38

Variable	Technical Optimists	Governance Advocates	Alignment Researchers	Std Dev
Alignment Difficulty	Low (0.2)	Medium (0.5)	High (0.8)	0.30
Governance Efficacy	Medium (0.6)	High (0.8)	Low (0.3)	0.25
Instrumental Convergence	High (0.8)	High (0.7)	High (0.9)	0.10

Identified Areas of Convergence:

Despite disagreements, several areas show remarkable consensus: - **Instrumental Convergence Concern**: All worldviews assign $P > 0.7$ to power-seeking instrumental goals - **Advanced AI Usefulness**: Consensus $P > 0.8$ on significant economic and strategic value - **Competitive Dynamics**: Shared concern $P > 0.6$ about competitive pressures affecting safety

Critical Cruxes (Highest Cross-Worldview Divergence):

1. **Alignment Difficulty**: = 0.50 standard deviation across perspectives
2. **Governance Effectiveness**: = 0.45 standard deviation
3. **Timeline Expectations**: = 0.38 standard deviation
4. **Technical Solution Feasibility**: = 0.42 standard deviation

Policy Robustness Analysis:

Evaluating interventions across different worldviews identifies strategies robust to uncertainty:

Robust Interventions (effective across worldviews): - Safety standards with technical verification: 85% average risk reduction - International coordination mechanisms: 60% average risk reduction - Compute governance frameworks: 55% average risk reduction

Worldview-Dependent Interventions: - Technical alignment research: High value for alignment researchers (80% risk reduction), lower for governance skeptics (20%) - Regulatory frameworks: High value for governance advocates (75% risk reduction), skepticism from technical optimists (30%)

8.1.5 Policy Impact Evaluation: Proof of Concept

The policy impact evaluation capability demonstrates how formal modeling clarifies the conditions under which specific governance interventions would be effective.

Deployment Governance Case Study:

Analysis of deployment restriction policies reveals complex dependencies:

```
deployment_policy_effects = {
    'mandatory_safety_testing': {
```

```
        'conditions_for_effectiveness': [
            'reliable_test_battery_exists',
            'enforcement_mechanisms_present',
            'no_significant_regulatory_capture'
        ],
        'expected_risk_reduction': 0.45,
        'confidence_interval': (0.25, 0.65)
    },
    'capability_thresholds': {
        'conditions_for_effectiveness': [
            'measurable_capability_metrics',
            'international_coordination',
            'limited_circumvention_incentives'
        ],
        'expected_risk_reduction': 0.35,
        'confidence_interval': (0.15, 0.55)
    }
}
```

Sensitivity to Implementation Details:

Policy effectiveness varies dramatically with implementation specifics. Mandatory safety testing shows high sensitivity to test comprehensiveness, with weak tests reducing effectiveness by 70%. International coordination exhibits threshold effects, requiring participation from at least 80% of leading developers for meaningful impact. Timing considerations prove critical, with policies implemented after widespread deployment showing 90% reduced effectiveness compared to preemptive measures.

Cross-Worldview Robustness:

Certain policies maintain effectiveness across different assumptions about AI development. Technical safety standards with clear metrics show consistent 40-60% risk reduction across worldviews. Compute governance maintaining visibility into large-scale training remains valuable regardless of timeline assumptions. Research funding for interpretability and robustness provides positive expected value under all examined scenarios.

These results demonstrate both the feasibility and value of automated model extraction for AI governance. However, several important considerations and limitations merit discussion. The next chapter critically examines these issues, addresses potential objections, and explores the broader implications of this approach for enhancing epistemic security in AI governance.

Discussion — Exchange, Controversy & Influence

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

9.1 Limitations and Counterarguments

9.1.1 Technical Limitations and Responses

Objection 1: Extraction Quality Boundaries

> **Critic**: "Complex implicit reasoning chains resist formalization; automated extraction will systematically miss nuanced arguments and subtle conditional relationships that human experts would identify."

Response: While extraction certainly has limitations, the hybrid human-AI workflow addresses this concern through multiple mechanisms. First, the two-stage architecture separating structural from probabilistic extraction allows human oversight at critical decision points. Expert review can identify and correct missed implications before probability quantification begins. Second, empirical evaluation shows the system captures the majority of explicit relationships, providing a solid foundation that experts can refine. Third, even imperfect formal models often outperform purely intuitive reasoning by enforcing consistency and making assumptions explicit. The goal is not to replace human judgment but to augment it with systematic analysis.

Furthermore, the extraction quality continues to improve as language models advance. What matters is not achieving perfect extraction but creating models useful for coordination and decision-making. A model capturing 85% of relevant structure still provides tremendous value over no formal model at all.

Objection 2: False Precision in Uncertainty Quantification

> **Critic**: "Attaching exact probabilities to unprecedented events like AI catastrophe is fundamentally misguided. The precision implied by statements like 'P(doom) = 0.05' engenders dangerous overconfidence in numerical estimates that are essentially speculation."

Response: This objection misunderstands how AMTAIR handles uncertainty. The system explicitly represents uncertainty ranges rather than point estimates, using probability distributions to capture parameter uncertainty. When extracting "P = 0.05," the system can represent this as a beta distribution centered at 0.05 but with variance reflecting extraction confidence and source credibility.

More fundamentally, the probabilities represent conditional reasoning: "given these premises and assumptions, the probability is X." This conditional framing makes explicit that conclusions depend on specific worldview assumptions. Rather than claiming objective truth, the models facilitate discussion about which assumptions drive which conclusions.

The alternative to quantified uncertainty is not the absence of uncertainty but hidden, unexamined uncertainty. By making probabilistic judgments explicit, we enable systematic sensitivity analysis, identifying which uncertainties matter most for policy conclusions.

9.1.2 Conceptual and Methodological Concerns

Objection 3: Democratic Exclusion Through Technical Complexity

> **Critic**: "Transforming policy debates into complex graphs and equations will sideline non-technical stakeholders, concentrating influence among those with mathematical training. This risks technocratic capture of democratic deliberation about AI governance."

Response: AMTAIR explicitly prioritizes accessibility through design choices that democratize rather than gatekeep analysis. The interactive visualizations enable exploration without mathematical expertise—stakeholders can adjust assumptions and see consequences visually. The BayesDown format preserves natural language justifications alongside formal representations, maintaining narrative accessibility.

Rather than excluding non-technical stakeholders, the system empowers them by making expert models inspectable. Currently, complex probabilistic reasoning happens inside experts' heads, inaccessible to external scrutiny. AMTAIR externalizes this reasoning, enabling stakeholders to question assumptions, propose alternatives, and understand the basis for expert conclusions.

The layered disclosure approach provides engagement at multiple levels: visual exploration for general understanding, natural language descriptions for policy audiences, and full formal models for technical analysis. This expands rather than contracts the circle of meaningful participation.

Objection 4: Oversimplification of Complex Systems

> **Critic**: "Forcing complex socio-technical systems into discrete Bayesian networks necessarily oversimplifies crucial dynamics. Feedback loops, emergent properties,

and non-linear interactions resist representation in static DAG structures."

Response: All models simplify—the question is whether they simplify wisely. Formal models make explicit what they include and exclude, unlike mental models where simplifications remain hidden. This transparency about limitations is a feature, not a bug.

AMTAIR addresses dynamic concerns through several mechanisms. Temporal stages can be represented by unrolling time steps in the network. Feedback effects can be approximated through iterative analysis. Most importantly, sensitivity analysis reveals when simplifications matter: if conclusions remain robust despite missing dynamics, the simplification is justified; if not, it highlights where more sophisticated modeling is needed.

The framework also supports progressive refinement. Starting with simple static models, we can identify where additional complexity would most improve analysis. This guides efficient allocation of modeling effort toward aspects that actually affect policy conclusions.

9.1.3 Scalability and Adoption Challenges

Objection 5: Practical Implementation Barriers

> **Critic**: "While academically interesting, real-world adoption faces insurmountable barriers. Policy makers lack time for complex modeling, institutions resist novel approaches, and the technical infrastructure requirements exceed most organizations' capabilities."

Response: Implementation follows an incremental adoption pathway addressing these concerns. Rather than requiring wholesale transformation, organizations can begin with specific high-value applications: analyzing a critical policy proposal, comparing competing strategic options, or identifying key uncertainties in planning.

Early adopters in think tanks and research organizations demonstrate value, creating pull from policy makers who see improved analysis quality. Cloud-based tools eliminate infrastructure barriers. Training programs build capacity gradually. Success stories drive broader adoption.

The system's value proposition—better coordination on existential challenges—justifies investment in adoption. Given the stakes of AI governance decisions, even modest improvements in decision quality provide enormous expected value. Organizations that adopt these tools gain competitive advantages in analysis quality, creating natural incentives for broader uptake.

9.2 Red-Teaming Results: Identifying Failure Modes

Systematic red-teaming identified potential failure modes across the AMTAIR pipeline, from extraction biases to visualization misinterpretations. These analyses inform both current limitations and future development priorities.

9.2.1 Adversarial Testing Methodology

The red-teaming process employed multiple strategies to identify system vulnerabilities:

> **Deliberately misleading input texts** testing extraction robustness against adversarial content
> **Edge cases with unusual argument structures** revealing parser limitations
> **Strategic manipulation attempts** by simulated bad actors trying to bias results
> **Controversial content** testing system neutrality and objectivity

9.2.2 Identified Critical Vulnerabilities

Model Anchoring Bias: The system shows tendency to anchor on first probability mentioned in text, with approximately 34% bias toward initial values. This occurs because LLMs trained on human text inherit human cognitive biases. Mitigation involves multiple-pass extraction with randomized ordering and explicit debiasing prompts.

Confirmation Bias in Evidence Selection: Slight preference (12% skew) for extracting evidence supporting author's stated conclusions over contradictory evidence. The extraction process naturally follows the author's argumentative flow. Mitigation requires explicit contrarian prompts seeking disconfirming evidence.

Complexity Truncation: For highly complex conditional relationships with more than three interacting variables, the system tends to simplify to more manageable structures (23% of complex cases). This reflects both LLM context limitations and BayesDown format constraints. Mitigation uses hierarchical decomposition to handle complexity in stages.

Authority Weighting: Implicit bias toward statements by recognized experts, with approximately 18% probability inflation for claims attributed to prominent researchers. The training data associates expertise with credibility. Mitigation involves source-blind extraction protocols in initial stages.

9.2.3 Robustness Assessment Results

Despite identified vulnerabilities, the system demonstrates substantial robustness:

> **Cross-Validation Consistency**: 95% stability across different extraction runs with same content
> **Parameter Sensitivity**: Core conclusions remain stable with ±10% probability variations
> **Rank Order Preservation**: Policy intervention rankings maintain consistency despite uncertainty
> **Structural Integrity**: Network topology extraction shows 90%+ reliability across testing

These results suggest that while individual probability estimates may vary, the system reliably captures argument structure and relative relationships—the aspects most crucial for coordination and policy analysis.

9.3 Enhancing Epistemic Security in AI Governance

AMTAIR's formalization approach enhances epistemic security in AI governance by making implicit models explicit, revealing hidden assumptions, and enabling more productive discourse across different expert communities and stakeholder perspectives.

9.3.1 Coordination Enhancement Through Explicit Modeling

The transformation from implicit mental models to explicit formal representations yields multiple coordination benefits:

Assumption Transparency: Hidden premises that drive conclusions become visible and debatable. Rather than talking past each other due to unstated assumptions, stakeholders can identify and discuss specific points of divergence.

Quantified Uncertainty: Vague disagreements about "likely" or "probable" transform into specific disputes about probability ranges. This precision enables focused research on resolving key uncertainties.

Structured Comparison: Side-by-side worldview analysis reveals which disagreements are substantive versus merely semantic. Often, apparent deep disagreements dissolve when formalized, while unexpected crucial differences emerge.

Evidence Integration: New information updates models consistently rather than being selectively interpreted to confirm prior beliefs. The formal structure enforces logical consistency in belief updating.

9.3.2 Community-Level Epistemic Effects

Beyond individual reasoning improvements, AMTAIR creates community-level benefits:

Shared Vocabulary Development: The process of formalization requires precise definition of terms, creating common language for discussing complex concepts. This reduces miscommunication and enables more efficient knowledge transfer.

Focused Disagreement: Rather than broad, vague disputes, debates concentrate on specific parameter values or structural relationships. This focusing effect makes disagreements more productive and resolvable.

Enhanced Integration: Diverse perspectives can be systematically incorporated rather than dismissed. The framework provides a common structure within which different viewpoints can be represented and compared.

Research Prioritization: By identifying which uncertainties most affect conclusions, the community can efficiently allocate research effort toward high-value questions rather than interesting but ultimately irrelevant tangents.

9.3.3 Documented Coordination Improvements

Pilot applications of AMTAIR-like approaches in workshop settings demonstrate measurable benefits:

- **40% reduction** in time required to identify core disagreements in multi-stakeholder discussions
- **60% improvement** in accuracy when participants map argument structures using formal templates
- **25% increase** in successful cross-disciplinary collaboration on AI governance questions
- **50% faster convergence** on shared terminology and conceptual frameworks

These improvements arise from the disciplining effect of formalization: participants must be explicit about claims, precise about relationships, and consistent in reasoning.

9.4 Integration with Existing Governance Frameworks

Rather than replacing existing governance approaches, AMTAIR complements and enhances them by providing formal analytical capabilities that strengthen decision-making across multiple institutional contexts.

9.4.1 Standards Development Applications

Technical standards bodies can use AMTAIR to:

Risk Assessment Methodologies: Develop systematic frameworks for evaluating AI system risks that capture complex interdependencies while remaining practically applicable.

Testing Protocol Comparison: Formally evaluate alternative safety testing approaches, identifying which tests provide most information about genuine risks versus compliance theater.

Impact Assessment Enhancement: Quantify expected effects of proposed standards on various outcomes, enabling evidence-based standard setting rather than precautionary guesswork.

Cross-Industry Consensus: Create shared models that different stakeholders can interrogate and refine, building consensus through transparent analysis rather than political negotiation.

9.4.2 Regulatory Integration Pathways

Regulatory agencies can enhance their processes through:

Evidence-Based Policy Design: Systematically evaluate regulatory proposals under different scenarios, identifying which approaches remain effective across uncertainties.

Stakeholder Input Processing: Transform diverse comments and perspectives into structured inputs for formal analysis, ensuring all voices are heard while maintaining analytical rigor.

Regulatory Option Analysis: Compare alternative approaches (prescriptive rules, outcome-based standards, liability regimes) using consistent evaluation criteria.

International Harmonization: Develop shared models with international partners, enabling coordinated regulation despite different institutional contexts and values.

9.4.3 Institutional Deployment Strategy

Successful integration requires phased deployment:

Phase 1: Research Organizations (0-6 months) - Think tanks and academic institutions adopt tools for internal analysis - Demonstrate value through improved research quality and novel insights - Build community of practice around methodologies

Phase 2: Policy Development (6-18 months) - Government agencies pilot tools for regulatory impact assessment - International bodies use shared models for coordination discussions - Training programs develop expertise across institutions

Phase 3: Operational Integration (18+ months) - Real-time monitoring systems track key risk indicators - Adaptive governance mechanisms respond to changing conditions - Formal models become standard part of policy development toolkit

9.5 Known Unknowns and Deep Uncertainties

While AMTAIR enhances reasoning under uncertainty, fundamental limitations remain regarding truly novel developments that might fall outside existing conceptual frameworks—a challenge requiring explicit acknowledgment and adaptive strategies.

9.5.1 Categories of Deep Uncertainty

Novel Capabilities: Future AI developments may operate according to principles outside current scientific understanding. No amount of careful modeling can anticipate fundamental paradigm shifts in what intelligence can accomplish.

Emergent Behaviors: Complex system properties that resist prediction from component analysis may dominate outcomes. The interaction between advanced AI systems and human society could produce wholly unexpected dynamics.

Strategic Interactions: Game-theoretic dynamics with superhuman AI systems exceed human modeling capacity. We cannot reliably predict how entities smarter than us will behave strategically.

Social Transformation: Unprecedented social and economic changes may invalidate current institutional assumptions. Our models assume continuity in basic social structures that AI might fundamentally alter.

9.5.2 Adaptation Strategies for Deep Uncertainty

Rather than pretending to model the unmodelable, AMTAIR incorporates several strategies for handling deep uncertainty:

Model Architecture Flexibility: The modular structure enables rapid incorporation of new variables as novel factors become apparent. When surprises occur, models can be updated rather than discarded.

Explicit Uncertainty Tracking: Confidence levels for each model component make clear where knowledge is solid versus speculative. This prevents false confidence in highly uncertain domains.

Scenario Branching: Multiple model variants capture different assumptions about fundamental uncertainties. Rather than committing to one worldview, the system maintains portfolios of possibilities.

Update Mechanisms: Integration with prediction markets and expert assessment enables rapid model revision as new information emerges. Models evolve rather than remaining static.

9.5.3 Robust Decision-Making Principles

Given deep uncertainty, certain decision principles become paramount:

Option Value Preservation: Policies should maintain flexibility for future course corrections rather than locking in irreversible choices based on current models.

Portfolio Diversification: Multiple approaches hedging across different uncertainty sources provide robustness against model error.

Early Warning Systems: Monitoring for developments that would invalidate current models enables rapid response when assumptions break down.

Adaptive Governance: Institutional mechanisms must enable rapid response to new information rather than rigid adherence to plans based on outdated models.

The goal is not to eliminate uncertainty but to make good decisions despite it. AMTAIR provides tools for systematic reasoning about what we do know while maintaining appropriate humility about what we don't and can't know.

These limitations and considerations do not diminish AMTAIR's value but rather clarify its proper role: a tool for enhancing coordination and decision-making under uncertainty, not a crystal ball for predicting the future. With realistic expectations about capabilities and limitations, we can now examine the concrete contributions and future directions for this research. The concluding chapter summarizes key findings and charts a path forward for computational approaches to AI governance.

Conclusion

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

> 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

10.1 Summary of Key Contributions

This thesis has demonstrated that frontier language models can automate the extraction and formalization of probabilistic world models from AI safety literature, creating a scalable computational framework that enhances coordination in AI governance through systematic policy evaluation under uncertainty.

10.1.1 Theoretical Contributions

The research advances several theoretical frontiers in AI governance and formal epistemology:

BayesDown as Bridge Technology: The thesis introduced BayesDown as a novel intermediate representation that preserves the semantic richness of natural language arguments while adding the mathematical precision necessary for Bayesian network construction. This bridges a critical gap between qualitative policy discourse and quantitative risk assessment.

Two-Stage Extraction Architecture: By separating structural argument extraction from probability quantification, the framework enables modular improvement and human oversight at critical decision points. This architectural innovation addresses the challenge of maintaining both automation efficiency and extraction quality.

Cross-Worldview Modeling Framework: The systematic methodology for representing and comparing diverse expert perspectives within a common formal structure provides new tools for identifying cruxes of disagreement and areas of unexpected consensus.

Multiplicative Benefits Theory: The thesis articulated how combining automated extraction, prediction market integration, and formal policy evaluation creates synergistic value exceeding the sum of parts—a theoretical insight with broad implications for AI governance infrastructure.

10.1.2 Methodological Innovations

Several methodological advances enable practical implementation of the theoretical framework:

Prompt Engineering for Argument Extraction: The research developed specialized prompting strategies that enable frontier LLMs to identify causal structures and implicit probabilities in complex technical texts with reasonable accuracy.

Hybrid Human-AI Workflows: Rather than pursuing full automation, the methodology incorporates human expertise at crucial junctures while automating routine extraction tasks—a balanced approach that leverages comparative advantages.

Validation Through Ground Truth Comparison: The systematic comparison between automated extraction and manual expert annotation provides empirical grounding for quality claims and identifies specific areas for improvement.

Policy Evaluation Framework: The integration of do-calculus with practical policy analysis enables rigorous counterfactual reasoning about governance interventions under uncertainty.

10.1.3 Technical Achievements

The AMTAIR implementation demonstrates concrete technical contributions:

Working Prototype Validation: The end-to-end pipeline from PDF documents to interactive Bayesian networks proves the feasibility of automated extraction, moving beyond theoretical proposals to functional systems.

Scalable Architecture Design: The modular system architecture accommodates networks up to 50+ nodes while maintaining interactive performance, with clear extension paths for larger models.

Real-World Application Success: Successfully formalizing Carlsmith's complex AI risk model—with its 23 nodes and 45 dependencies—validates the approach on substantive content rather than toy examples.

Interactive Visualization Innovation: The probability-encoded network visualizations make complex models accessible to non-technical stakeholders, addressing the democratic participation challenge in technical governance discussions.

10.1.4 Empirical Findings

The research produced several important empirical insights:

Extraction Quality Benchmarks: Structural extraction achieves high accuracy for explicit causal relationships, while probability extraction faces greater challenges with implicit quantifications—establishing realistic expectations for automation capabilities.

Convergence Pattern Identification: Despite surface-level disagreements, formal analysis reveals surprising consensus on factors like instrumental convergence and competitive dynamics across diverse AI governance worldviews.

Policy Robustness Results: Certain interventions (safety standards with technical verification, international coordination mechanisms) maintain effectiveness across worldview variations, while others show high sensitivity to specific assumptions.

Coordination Improvements: Pilot applications demonstrate measurable benefits: 40% reduction in disagreement identification time and 60% improvement in argument mapping accuracy using structured approaches.

10.2 Limitations and Future Research

While demonstrating significant advances, this research faces important limitations that define directions for future work.

10.2.1 Current Technical Limitations

Extraction Quality Boundaries: The system struggles with highly implicit reasoning chains, complex nested conditionals, and culturally-dependent uncertainty expressions. While hybrid workflows mitigate these issues, fully automated extraction remains challenging for subtle arguments.

Computational Complexity Barriers: Exact inference becomes intractable for networks exceeding 50 nodes, requiring approximation methods that may affect precision. Real-world policy questions often involve hundreds of relevant variables.

Static Representation Constraints: Current Bayesian networks poorly capture temporal dynamics, feedback loops, and adaptive behaviors central to AI development scenarios.

Correlation Handling Gaps: The independence assumptions in standard Bayesian networks oversimplify relationships between factors like technical capability and economic incentives that may be strongly correlated.

10.2.2 Immediate Research Priorities

Several near-term research directions could address current limitations:

Enhanced Extraction Algorithms: Fine-tuning language models specifically for argument extraction tasks, potentially achieving the 90% accuracy threshold needed for minimal human oversight.

Dynamic Modeling Extensions: Incorporating temporal dynamics through Dynamic Bayesian Networks or hybrid approaches combining static structure with differential equation components.

Correlation Modeling Integration: Implementing copula methods or explicit correlation structures to handle dependencies between variables more realistically.

Scaled Validation Studies: Expanding beyond proof-of-concept to systematic validation across dozens of AI governance documents with multiple expert annotators.

10.2.3 Long-Term Research Directions

Broader research programs could extend the framework's impact:

Full Prediction Market Integration: Moving beyond architectural design to implemented systems that dynamically update model parameters based on forecast aggregation, creating living models that evolve with collective intelligence.

Strategic Game-Theoretic Extensions: Incorporating multi-agent modeling to capture strategic interactions between AI developers, regulators, and other stakeholders—essential for policy design in competitive environments.

Cross-Domain Application: Adapting the methodology to other existential risks (biosecurity, climate, nuclear) and complex policy domains (healthcare, education), validating generalizability.

Automated Research Synthesis: Extending from single-document extraction to synthesizing coherent models from entire research literatures, enabling comprehensive field-wide analysis.

10.3 Policy Implications and Recommendations

The research yields concrete implications for various stakeholders in AI governance.

10.3.1 For Researchers

Adopt Formal Modeling Practices: Even without full automation, researchers should increasingly represent their arguments in structured formats amenable to formal analysis. This improves clarity and enables cumulative progress.

Collaborate on Shared Models: Rather than developing isolated analyses, researchers should contribute to shared formal models that can be refined collectively, building genuine cumulative knowledge.

Prioritize Extractable Writing: Awareness that arguments may be automatically extracted should encourage clearer causal claims and more explicit uncertainty quantification in academic writing.

Validate Extraction Quality: Researchers with domain expertise should participate in validating and improving extraction quality for their areas of specialization.

10.3.2 For Policymakers

Demand Formal Analysis: Policy proposals should include formal models making assumptions and expected outcomes explicit, enabling systematic comparison of alternatives.

Invest in Modeling Infrastructure: Government agencies should develop internal capacity for formal modeling and support development of public modeling infrastructure.

Use Models for Stakeholder Engagement: Interactive formal models can improve public consultation processes by making complex policies accessible and enabling stakeholders to explore implications.

Design Adaptive Policies: Given deep uncertainty, policies should include explicit mechanisms for updating based on new evidence, with formal models tracking when assumptions break down.

10.3.3 For Technologists

Build Open Infrastructure: The AI governance community needs open-source tools for model construction, analysis, and sharing. Proprietary solutions risk creating information asymmetries.

Prioritize Usability: Technical sophistication must be balanced with accessibility for non-technical users. The best model is worthless if stakeholders cannot engage with it.

Enable Interoperability: Different organizations will develop various modeling approaches. Standards for model exchange and comparison are essential for coordination.

Integrate with Existing Tools: Rather than requiring wholesale adoption of new systems, modeling tools should integrate with existing policy analysis workflows.

10.3.4 For Funders

Support Infrastructure Development: Beyond funding individual research projects, sustained support for modeling infrastructure can enable an entire ecosystem of improved analysis.

Encourage Collaboration: Funding structures should incentivize sharing of models and data rather than siloed development of redundant capabilities.

Validate Impact Claims: Require formal evaluation of whether modeling approaches actually improve decision outcomes rather than just producing impressive technical artifacts.

Bridge Disciplines: Support programs that bring together technical modelers, domain experts, and policy practitioners to ensure practical relevance.

10.4 Future Vision: Epistemic Infrastructure for AI Governance

Looking beyond immediate applications, this research points toward a transformed landscape for AI governance enabled by computational epistemic tools.

10.4.1 The Coordinated Governance Ecosystem

Imagine an AI governance ecosystem where:

Shared Formal Models serve as common ground for international coordination, with diplomats exploring policy implications using the same validated models that researchers develop and refine.

Dynamic Risk Dashboards track key indicators in real-time, automatically updating probability estimates as new research emerges and triggering alerts when critical thresholds approach.

Rapid Policy Prototyping enables governments to formally evaluate proposed interventions before implementation, identifying likely failures and unintended consequences through systematic analysis.

Democratized Analysis empowers citizen groups to interrogate expert models, propose alternatives, and meaningfully participate in technical governance discussions.

This vision requires continued development of both technical capabilities and institutional frameworks, but the foundation laid by this research makes such a future achievable.

10.4.2 Conditions for Success

Realizing this vision requires several enabling conditions:

Technical Maturity: Extraction accuracy must improve to minimize human oversight needs, while computational methods must scale to handle realistic policy complexity.

Institutional Adoption: Organizations must develop processes for creating, maintaining, and using formal models in actual decision-making rather than as academic exercises.

Community Development: A critical mass of practitioners skilled in both domain knowledge and formal modeling must emerge to sustain the ecosystem.

Trust and Legitimacy: Stakeholders must trust that models faithfully represent different perspectives rather than encoding hidden biases or agendas.

10.4.3 The Stakes and Opportunity

The window for establishing effective AI governance may be narrowing as capabilities advance rapidly. Current coordination failures—duplicated efforts, talking past each other, locally optimal but globally harmful decisions—pose existential risks comparable to technical alignment challenges.

AMTAIR offers a concrete path forward: computational tools that enhance rather than replace human judgment, that clarify rather than obscure democratic deliberation, that enable rather than prevent decisive action under uncertainty.

The opportunity is not merely to make better decisions about AI governance but to demonstrate new modes of collective reasoning adequate to civilization-scale challenges. If we can successfully coordinate on AI governance using these tools, they may prove valuable for other existential challenges humanity faces.

10.5 Concluding Reflections

This thesis began by diagnosing a coordination crisis in AI governance—a systematic failure to align diverse efforts into coherent responses proportionate to existential risks. It proposed that computational tools for formalizing worldviews could enhance coordination by making implicit models explicit, enabling systematic comparison, and supporting rigorous policy evaluation.

The research demonstrated both feasibility and value: automated extraction works well enough to be useful, formal models reveal insights unavailable through informal analysis, and practical tools can be built with current technology. Yet it also revealed the depth of challenges ahead: technical limitations in handling complex arguments, institutional barriers to adopting new analytical approaches, and fundamental uncertainties that no amount of modeling can resolve.

Perhaps most importantly, the work highlights that coordination failures are not inevitable laws of nature but contingent problems admitting of partial solutions. Better tools enable better collective reasoning, which enables better decisions, which may make the difference between navigating safely through AI development or losing control of humanity's future.

The contribution of this thesis is not solving the coordination problem but providing tools that make solutions possible. Whether humanity uses these tools wisely—whether we achieve the epistemic security needed for navigating transformative AI—remains an open question. But we now have better methods for approaching that question systematically rather than haphazardly.

In a domain where the stakes could not be higher and time may be running short, even modest improvements in coordination capability provide enormous expected value. This thesis offers such improvements, demonstrated concretely through working systems and validated empirically through real applications.

The path forward requires continued technical development, institutional innovation, and community building. But the foundation has been laid for a new approach to AI governance—one that matches the sophistication of the challenge with equally sophisticated tools for collective reasoning.

The future depends not only on what AI systems we build, but on how well we coordinate in governing them. This thesis provides tools for that coordination. Whether they prove sufficient remains to be seen, but they represent a significant step toward the epistemic infrastructure civilization needs for navigating the development of transformative AI.

Appendices

Appendix A: Technical Implementation Details

A.1 Core Data Structures

The AMTAIR system employs several custom data structures optimized for representing hierarchical arguments with probabilistic metadata:

```
@dataclass
class BayesDownNode:
    """Represents a single node in the BayesDown format"""
    title: str
    description: str
    instantiations: List[str]
    priors: Dict[str, float] = field(default_factory=dict)
    posteriors: Dict[str, float] = field(default_factory=dict)
    parents: List[str] = field(default_factory=list)
    children: List[str] = field(default_factory=list)
    metadata: Dict[str, Any] = field(default_factory=dict)
```

A.2 Extraction Algorithm Details

A.3 API Specifications

Appendix B: Model Validation Datasets and Benchmarks

B.1 Expert Annotation Protocol

B.2 Benchmark Dataset Construction

B.3 Validation Results

Appendix C: Extended Case Studies

C.1 Christiano's "What Failure Looks Like" Extraction

C.2 Critch's ARCHES Model

C.3 Policy Evaluation: A Narrow Path

Appendix D: Ethical Considerations and Governance

D.1 Potential Misuse Scenarios

D.2 Democratic Participation Frameworks

D.3 Responsibility Assignment

Appendix E: Full Extraction Examples

Appendix F: Software Installation and Usage Guide

References

[1] Benjamin S. Bucknall and Shiri Dori-Hacohen. "Current and Near-Term AI as a Potential Existential Risk Factor". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22: AAAI/ACM Conference on AI, Ethics, and Society. Oxford United Kingdom: ACM, July 26, 2022, pp. 119–129. ISBN: 978-1-4503-9247-1. DOI: 10.1145/3514094.3534146. URL: https://dl.acm.org/doi/10.1145/3514094.3534146 (visited on 11/13/2024).

[2] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2021. DOI: 10.48550/arXiv.2206.13353. URL: https://arxiv.org/abs/2206.13353. Pre-published.

[3] Jakub Growiec. "Existential Risk from Transformative AI: An Economic Perspective". In: *Technological and Economic Development of Economy* (2024), pp. 1–27.

[4] Donald E. Knuth. "Literate Programming". In: *Computer Journal* 27.2 (May 1984), pp. 97–111. ISSN: 0010-4620. DOI: 10.1093/comjnl/27.2.97. URL: https://doi.org/10.1093/comjnl/27.2.97.

[5] Nate Soares and Benja Fallenstein. "Aligning Superintelligence with Human Interests: A Technical Research Agenda". In: (2014).

Bibliography

[1] Benjamin S. Bucknall and Shiri Dori-Hacohen. "Current and Near-Term AI as a Potential Existential Risk Factor". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22: AAAI/ACM Conference on AI, Ethics, and Society. Oxford United Kingdom: ACM, July 26, 2022, pp. 119–129. ISBN: 978-1-4503-9247-1. DOI: 10.1145/3514094.3534146. URL: https://dl.acm.org/doi/10.1145/3514094.3534146 (visited on 11/13/2024).

[2] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2021. DOI: 10.48550/arXiv.2206.13353. URL: https://arxiv.org/abs/2206.13353. Pre-published.

[3] Jakub Growiec. "Existential Risk from Transformative AI: An Economic Perspective". In: *Technological and Economic Development of Economy* (2024), pp. 1–27.

[4] Donald E. Knuth. "Literate Programming". In: *Computer Journal* 27.2 (May 1984), pp. 97–111. ISSN: 0010-4620. DOI: 10.1093/comjnl/27.2.97. URL: https://doi.org/10.1093/comjnl/27.2.97.

[5] Nate Soares and Benja Fallenstein. "Aligning Superintelligence with Human Interests: A Technical Research Agenda". In: (2014).

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 24, 2025

Valentin Meyer