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

**Preface** **1**

**Quarto Syntax** **3**

Main Formatting . . . 3
Html Comments . . . 3
Syntax for Tasks . . . 3
Tasks with ToDo Tree . . . 3
Task Syntax Examples . . . 4
Verbatim Code Formatting . . . 5
Code Block formatting . . . 5
Blockquote Formatting . . . 5
Tables . . . 5
Headings & Potential Headings in Standard Markdown formatting ('##') . . . 7
Heading 3 . . . 7
Text Formatting Options . . . 7
Lists . . . 7
Math . . . 7
Footnotes . . . 8
Callouts . . . 8
Links . . . 9
Page Breaks . . . 9
Including Code . . . 12
In-Line LaTeX . . . 12
In-Line HTML . . . 12
Reference or Embed Code from .ipynb files . . . 12
Diagrams . . . 15
Narrative citations (author as subject) . . . 16
Parenthetical citations (supporting reference) . . . 16
Author-only citation (when discussing the person) . . . 16
Year-only citation (when author already mentioned) . . . 16
Page-specific references . . . 16
Multiple works, different pages . . . 17

Section Cross-References . . . 17
Section Numbers . . . 17
Pages in Landscape . . . 17

**Abstract** **19**

**Prefatory Apparatus: Frontmatter** **21**
Illustrations and Terminology — Quick References . . . 21
**Acknowledgments** . . . 21
List of Graphics & Figures . . . 21
List of Abbreviations . . . 21

**Preface** **25**
Acknowledgments . . . 25

**Table of Contents** **27**

**Abstract** **29**

**List of Figures** **31**

**List of Tables** **33**

**List of Abbreviations** **35**

**1. Introduction: The Coordination Crisis in AI Governance** **37**
1.1 Opening Scenario: The Policymaker's Dilemma . . . 37
1.2 The Coordination Crisis in AI Governance . . . 37
1.2.1 Safety Gaps from Misaligned Efforts . . . 38
1.2.2 Resource Misallocation . . . 38
1.2.3 Negative-Sum Dynamics . . . 38
1.3 Historical Parallels and Temporal Urgency . . . 38
1.4 Research Question and Scope . . . 39
1.5 The Multiplicative Benefits Framework . . . 40
1.5.1 Automated Worldview Extraction . . . 40
1.5.2 Live Data Integration . . . 40
1.5.3 Formal Policy Evaluation . . . 41
1.5.4 The Synergy . . . 41
1.6 Thesis Structure and Roadmap . . . 41

**2. Context and Theoretical Foundations** **43**
2.1 AI Existential Risk: The Carlsmith Model . . . 43
2.1.1 Six-Premise Decomposition . . . 43
2.1.2 Why Carlsmith Exemplifies Formalizable Arguments . . . 44
2.2 The Epistemic Challenge of Policy Evaluation . . . 44
2.2.1 Unique Characteristics of AI Governance . . . 44

2.2.2 Limitations of Traditional Approaches . . . . . 45
2.3 Bayesian Networks as Knowledge Representation . . . . . 45
2.3.1 Mathematical Foundations . . . . . 45
2.3.2 The Rain-Sprinkler-Grass Example . . . . . 46
2.3.3 Advantages for AI Risk Modeling . . . . . 46
2.4 Argument Mapping and Formal Representations . . . . . 47
2.4.1 From Natural Language to Structure . . . . . 47
2.4.2 ArgDown: Structured Argument Notation . . . . . 47
2.4.3 BayesDown: The Bridge to Bayesian Networks . . . . . 47
2.5 The MTAIR Framework: Achievements and Limitations . . . . . 48
2.5.1 MTAIR's Approach . . . . . 48
2.5.2 Key Achievements . . . . . 48
2.5.3 Fundamental Limitations . . . . . 49
2.5.4 The Automation Opportunity . . . . . 49
2.6 Requirements for Coordination Infrastructure . . . . . 49
2.6.1 Scalability . . . . . 50
2.6.2 Accessibility . . . . . 50
2.6.3 Epistemic Virtues . . . . . 50
2.6.4 Integration Capabilities . . . . . 50
2.6.5 Robustness Properties . . . . . 50

**3. AMTAIR: Design and Implementation 53**
3.1 System Architecture Overview . . . . . 53
3.1.1 Five-Stage Pipeline . . . . . 53
3.1.2 Component Architecture . . . . . 53
3.2 The Two-Stage Extraction Process . . . . . 54
3.2.1 Stage 1: Structural Extraction (ArgDown) . . . . . 54
3.2.2 Stage 2: Probability Integration (BayesDown) . . . . . 54
3.2.3 Why Two Stages? . . . . . 55
3.3 Implementation Details . . . . . 55
3.3.1 Technology Stack . . . . . 55
3.3.2 Key Algorithms . . . . . 55
3.3.3 Performance Characteristics . . . . . 56
3.4 Case Study: Rain-Sprinkler-Grass . . . . . 56
3.4.1 Input Representation . . . . . 56
3.4.2 Processing Steps . . . . . 57
3.4.3 Results . . . . . 57
3.5 Case Study: Carlsmith's Power-Seeking AI Model . . . . . 57
3.5.1 Model Complexity . . . . . 57
3.5.2 Extraction Results . . . . . 57
3.5.3 Validation Against Original . . . . . 58
3.5.4 Insights from Formalization . . . . . 58
3.6 Validation Methodology . . . . . 59

3.6.1 Ground Truth Construction . . . . . 59
3.6.2 Evaluation Metrics . . . . . 59
3.6.3 Results Summary . . . . . 59
3.6.4 Error Analysis . . . . . 60
3.7 Policy Evaluation Capabilities . . . . . 60
3.7.1 Intervention Representation . . . . . 60
3.7.2 Example: Deployment Governance . . . . . 60
3.7.3 Robustness Analysis . . . . . 60
3.8 Interactive Visualization Design . . . . . 61
3.8.1 Visual Encoding Strategy . . . . . 61
3.8.2 Progressive Disclosure . . . . . 61
3.8.3 User Interface Elements . . . . . 61
3.9 Integration with Prediction Markets . . . . . 61
3.9.1 Design for Integration . . . . . 61
3.9.2 Challenges and Opportunities . . . . . 62
3.10 Computational Considerations . . . . . 62
3.10.1 Exact vs. Approximate Inference . . . . . 62
3.10.2 Scaling Strategies . . . . . 63
3.11 Summary of Technical Achievements . . . . . 63

**4. Discussion: Implications and Limitations 65**
4.1 Technical Limitations and Responses . . . . . 65
4.1.1 Objection 1: Extraction Quality Boundaries . . . . . 65
4.1.2 Objection 2: False Precision in Uncertainty . . . . . 66
4.1.3 Objection 3: Correlation Complexity . . . . . 66
4.2 Conceptual and Methodological Concerns . . . . . 67
4.2.1 Objection 4: Democratic Exclusion . . . . . 67
4.2.2 Objection 5: Oversimplification of Complex Systems . . . . . 68
4.3 Red-Teaming Results . . . . . 68
4.3.1 Adversarial Extraction Attempts . . . . . 68
4.3.2 Robustness Findings . . . . . 69
4.3.3 Implications for Deployment . . . . . 69
4.4 Enhancing Epistemic Security . . . . . 70
4.4.1 Making Models Inspectable . . . . . 70
4.4.2 Revealing Convergence and Divergence . . . . . 70
4.4.3 Improving Collective Reasoning . . . . . 70
4.5 Scaling Challenges and Opportunities . . . . . 71
4.5.1 Technical Scaling . . . . . 71
4.5.2 Social and Institutional Scaling . . . . . 71
4.5.3 Opportunities for Impact . . . . . 72
4.6 Integration with Governance Frameworks . . . . . 72
4.6.1 Standards Development . . . . . 72
4.6.2 Regulatory Design . . . . . 72

4.6.3 International Coordination . . . . . 73
4.6.4 Organizational Decision-Making . . . . . 73
4.7 Future Research Directions . . . . . 73
4.7.1 Technical Enhancements . . . . . 73
4.7.2 Methodological Extensions . . . . . 73
4.7.3 Application Domains . . . . . 74
4.7.4 Ecosystem Development . . . . . 74

**5. Conclusion: Toward Coordinated AI Governance 75**
5.1 Summary of Key Contributions . . . . . 75
5.1.1 Theoretical Contributions . . . . . 75
5.1.2 Methodological Innovations . . . . . 75
5.1.3 Technical Achievements . . . . . 76
5.1.4 Empirical Findings . . . . . 76
5.2 Limitations and Honest Assessment . . . . . 76
5.2.1 Technical Constraints . . . . . 76
5.2.2 Conceptual Limitations . . . . . 77
5.2.3 Practical Constraints . . . . . 77
5.3 Implications for AI Governance . . . . . 77
5.3.1 Near-Term Applications . . . . . 77
5.3.2 Medium-Term Transformation . . . . . 78
5.3.3 Long-Term Vision . . . . . 78
5.4 Recommendations for Stakeholders . . . . . 79
5.4.1 For Researchers . . . . . 79
5.4.2 For Policymakers . . . . . 79
5.4.3 For Technologists . . . . . 79
5.4.4 For Funders . . . . . 79
5.5 Future Research Agenda . . . . . 80
5.5.1 Technical Priorities . . . . . 80
5.5.2 Methodological Development . . . . . 80
5.5.3 Application Expansion . . . . . 81
5.6 Closing Reflections . . . . . 81

**References 83**

**Appendices 85**
Appendix A: Technical Implementation Details . . . . . 85
Appendix B: Validation Datasets and Procedures . . . . . 85
Appendix C: Extended Case Studies . . . . . 85
Appendix D: BayesDown Syntax Specification . . . . . 85
Appendix E: Prompt Engineering Details . . . . . 85
Appendix F: User Guide . . . . . 85
Appendix G: Jupyter Notebook Implementation . . . . . 85

# List of Figures

1 Five-step AMTAIR automation pipeline from PDFs to Bayesian networks . . . . 10
2 Short 2 caption . . . . . 10
3 . . . . . 12

# List of Tables

1 Demonstration of pipe table syntax . . . . . 5
2 My Caption 1 . . . . . 5
3 Main Caption . . . . . 6
4 Sample grid table. . . . . 6
5 Performance benchmarks for different network sizes . . . . . 56
6 Carlsmith model extraction validation results . . . . . 58
7 System validation results across components . . . . . 59

# 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<br>• bright color |
| Oranges | $2.10 | • cures scurvy<br>• 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
```

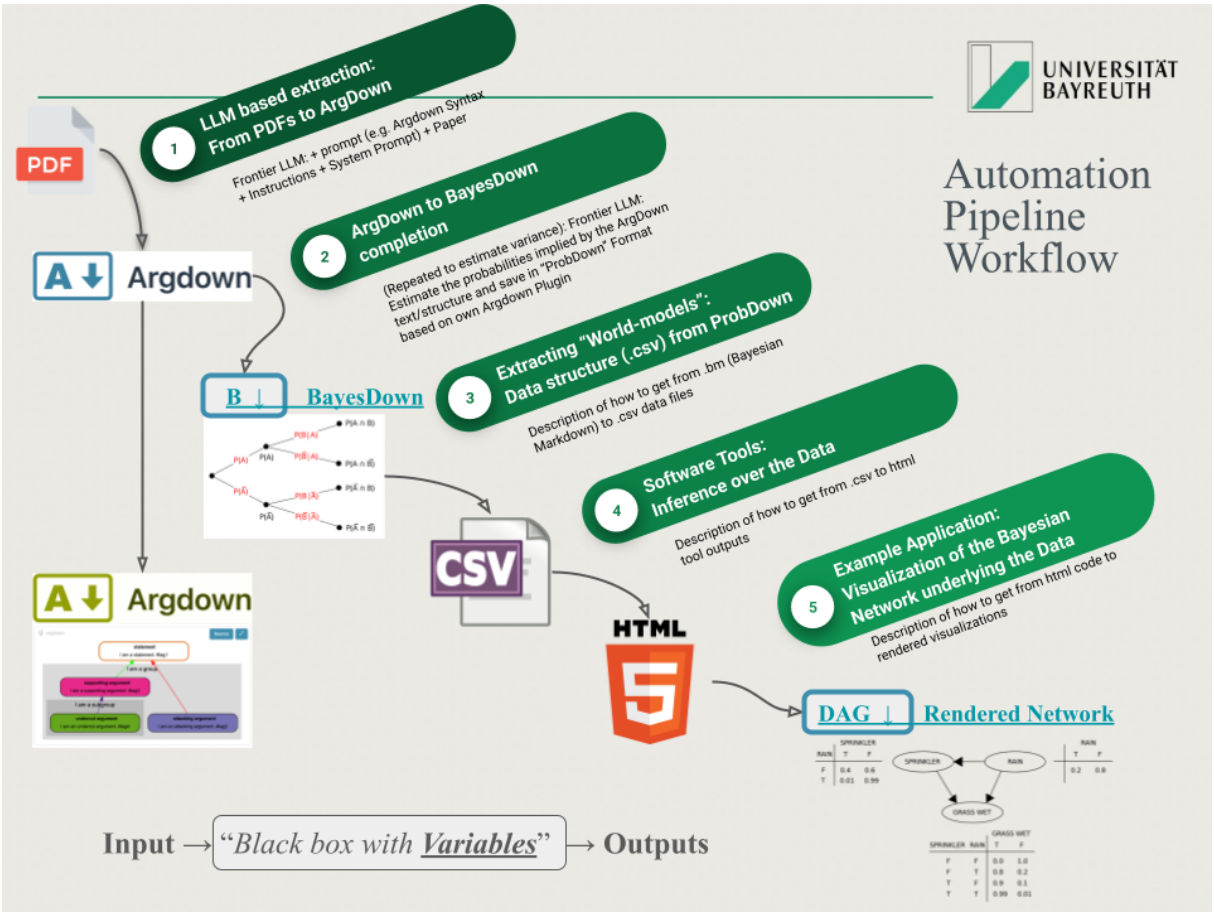



Figure 1: AMTAIR Automation Pipeline from

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:

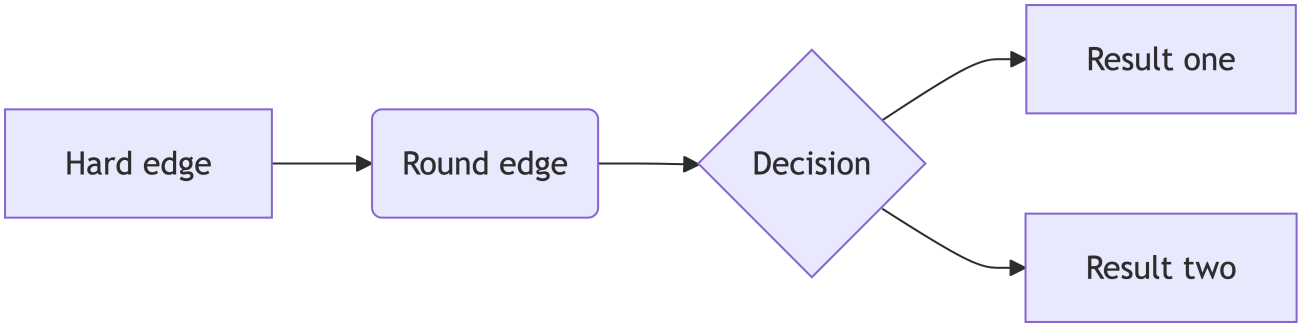
Full Notebooks are embedded in the Appendix through the _quarto.yml file with:

## 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 [3]

[3] and [2]

Blah Blah [see 2, pp. 33–35, also 1, chap. 1]

Blah Blah [2, 33–35, 38-39 and passim]

Blah Blah [1, 2].

Growiec says blah [1]

### Narrative citations (author as subject)

Soares and Fallenstein [3] argues that AI alignment requires…

### Parenthetical citations (supporting reference)

Recent work supports this view [3, 2].

### Author-only citation (when discussing the person)

As [3] demonstrates in their analysis…

### Year-only citation (when author already mentioned)

Soares [3] later revised this position.

### Page-specific references

The key insight appears in [3, pp. 45–67].

### Multiple works, different pages

This view is supported [3, p. 23, 2, 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 **?@sec-carlsmith-model**

`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 **?@sec-carlsmith-model**

# 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


> Figure 1.1: The coordination crisis in AI governance - visualization of fragmentation

> Figure 2.1: The Carlsmith model - DAG representation

> Figure 3.1: Research design overview - workflow diagram

> Figure 3.2: From natural language to BayesDown - transformation process

> Figure 4.1: ARPA system architecture - component diagram

> Figure 4.2: Visualization of Rain-Sprinkler-Grass_Wet Bayesian network - screenshot

> Figure 5.1: Extraction quality metrics - comparative chart

> Figure 5.2: Comparative analysis of AI governance worldviews - network visualization


## 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)**

# Preface

This thesis represents the culmination of interdisciplinary research at the intersection of AI safety, formal epistemology, and computational social science. The work emerged from recognizing a fundamental challenge in AI governance: while investment in AI safety research has grown exponentially, coordination between different stakeholder communities remains fragmented, potentially increasing existential risk through misaligned efforts.

The journey from initial concept to working implementation involved iterative refinement based on feedback from advisors, domain experts, and potential users. What began as a technical exercise in automated extraction evolved into a broader framework for enhancing epistemic security in one of humanity's most critical coordination challenges.

I hope this work contributes to building the intellectual and technical infrastructure necessary for humanity to navigate the transition to transformative AI safely. The tools and frameworks presented here are offered in the spirit of collaborative problem-solving, recognizing that the challenges we face require unprecedented cooperation across disciplines, institutions, and worldviews.

## Acknowledgments

I thank my supervisor Dr. Timo Speith for guidance throughout this project, the MTAIR team for pioneering the manual approach that inspired automation, and the AI safety community for creating the rich literature that made this work possible. Special recognition goes to technical advisors who provided implementation feedback and domain experts who validated extraction results.

Add specific names of: - MTAIR team members - Technical advisors - Domain experts who participated in validation - Funding sources if applicable

This research was conducted with support from [funding sources] and benefited from computational resources provided by [institutions]. Any errors or limitations remain my own responsibility.

# Table of Contents

# Abstract

# List of Figures

Auto-generated by Quarto with proper figure captions and labels

# List of Tables

Auto-generated by Quarto with proper table captions and labels

# List of Abbreviations

AI - Artificial Intelligence
AGI - Artificial General Intelligence
AMTAIR - Automating Transformative AI Risk Modeling
API - Application Programming Interface
APS - Advanced, Planning, Strategic (AI systems)
BN - Bayesian Network
CPT - Conditional Probability Table
DAG - Directed Acyclic Graph
LLM - Large Language Model
MTAIR - Modeling Transformative AI Risks
TAI - Transformative Artificial Intelligence

# 1. Introduction: The Coordination Crisis in AI Governance

## 1.1 Opening Scenario: The Policymaker's Dilemma

Imagine a senior policy advisor preparing recommendations for AI governance legislation. On her desk lie a dozen reports from leading AI safety researchers, each painting a different picture of the risks ahead. One argues that misaligned AI could pose existential risks within the decade, citing complex technical arguments about instrumental convergence and orthogonality. Another suggests these concerns are overblown, emphasizing uncertainty and the strength of existing institutions. A third proposes specific technical standards but acknowledges deep uncertainty about their effectiveness.

Each report seems compelling in isolation, written by credentialed experts with sophisticated arguments. Yet they reach dramatically different conclusions about both the magnitude of risk and appropriate interventions. The technical arguments involve unfamiliar concepts—mesa-optimization, corrigibility, capability amplification—expressed through different frameworks and implicit assumptions. Time is limited, stakes are high, and the legislation could shape humanity's trajectory for decades.

This scenario plays out daily across government offices, corporate boardrooms, and research institutions worldwide. It exemplifies what I term the "coordination crisis" in AI governance: despite unprecedented attention and resources directed toward AI safety, we lack the epistemic infrastructure to synthesize diverse expert knowledge into actionable governance strategies.

## 1.2 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 emerging multimodal systems—humanity faces a governance challenge unlike any in history. The task of ensuring increasingly powerful AI systems remain aligned with human values and beneficial to our long-term flourishing grows more urgent with each capability breakthrough. This challenge becomes particularly acute when considering transformative AI systems that could drastically alter civilization's trajectory, potentially including existential risks from misaligned systems pursuing objectives counter to human welfare.

The current state of AI governance presents a striking paradox. On one hand, we witness extraordinary mobilization: billions in research funding, proliferating safety initiatives, major tech companies establishing alignment teams, and governments worldwide developing AI strategies. The Asilomar AI Principles garnered thousands of signatures, the EU advances comprehensive AI regulation, and technical researchers produce increasingly sophisticated work on alignment, interpretability, and robustness.

Yet alongside this activity, we observe systematic coordination failures that may prove catastrophic. Technical safety researchers develop sophisticated alignment techniques without clear implementation pathways. Policy specialists craft regulatory frameworks lacking technical grounding to ensure practical efficacy. Ethicists articulate normative principles that lack operational specificity. Strategy researchers identify critical uncertainties but struggle to translate these into actionable guidance. International bodies convene without shared frameworks for assessing interventions.

### 1.2.1 Safety Gaps from Misaligned Efforts

When different communities operate with incompatible frameworks, critical risks fall through the cracks. Technical researchers may solve alignment problems under assumptions that policymakers' decisions invalidate. Regulations optimized for current systems may inadvertently incentivize dangerous development patterns. Without shared models of the risk landscape, our collective efforts resemble the parable of blind men describing an elephant—each accurate within their domain but missing the complete picture.

### 1.2.2 Resource Misallocation

The AI safety community duplicates efforts while leaving critical areas underexplored. Multiple teams independently develop similar frameworks without building on each other's work. Funders struggle to identify high-impact opportunities across technical and governance domains. Talent flows toward well-publicized approaches while neglected strategies remain understaffed. This misallocation becomes more costly as the window for establishing effective governance narrows.

### 1.2.3 Negative-Sum Dynamics

Perhaps most concerning, uncoordinated interventions can actively increase risk. Safety standards that advantage established players may accelerate risky development elsewhere. Partial transparency requirements might enable capability advances without commensurate safety improvements. International agreements lacking shared technical understanding may lock in dangerous practices. Without coordination, our cure risks becoming worse than the disease.

## 1.3 Historical Parallels and Temporal Urgency

History offers instructive parallels. The nuclear age began with scientists racing to understand and control forces that could destroy civilization. Early coordination failures—competing national programs, scientist-military tensions, public-expert divides—nearly led to catastrophe

multiple times. Only through developing shared frameworks (deterrence theory), institutions (IAEA), and communication channels (hotlines, treaties) did humanity navigate the nuclear precipice.

Yet AI presents unique coordination challenges that compress our response timeline:

**Accelerating Development**: Unlike nuclear weapons requiring massive infrastructure, AI development proceeds in corporate labs and academic departments worldwide. Capability improvements come through algorithmic insights and computational scale, both advancing exponentially.

**Dual-Use Ubiquity**: Every AI advance potentially contributes to both beneficial applications and catastrophic risks. The same language model architectures enabling scientific breakthroughs could facilitate dangerous manipulation or deception at scale.

**Comprehension Barriers**: Nuclear risks were viscerally understandable—cities vaporized, radiation sickness, nuclear winter. AI risks involve abstract concepts like optimization processes, goal misspecification, and emergent capabilities that resist intuitive understanding.

**Governance Lag**: Traditional governance mechanisms—legislation, international treaties, professional standards—operate on timescales of years to decades. AI capabilities advance on timescales of months to years, creating an ever-widening capability-governance gap.

## 1.4 Research Question and Scope

This thesis addresses a specific dimension of the coordination challenge by investigating:

**How can computational approaches formalize the worldviews and arguments underlying AI safety discourse, transforming qualitative disagreements into quantitative models suitable for rigorous policy evaluation?**

More specifically, I explore whether frontier AI technologies can be utilized to automate the modeling of transformative AI risks, enabling robust prediction of policy impacts across diverse worldviews. This investigation breaks down into several components:

> **Computational Formalization**: Using automated extraction and formal representation to make implicit models explicit
> **Worldview Representation**: Capturing different perspectives on AI risk in comparable frameworks
> **Argument Transformation**: Converting natural language arguments into structured Bayesian networks
> **Policy Evaluation**: Assessing intervention impacts through formal counterfactual analysis

The scope encompasses both theoretical development and practical implementation. Theoretically, I develop a framework for representing diverse perspectives on AI risk in a common formal language. Practically, I implement this framework in a computational system—the AI Risk Pathway Analyzer (ARPA)—that enables interactive exploration of how policy interventions might

alter existential risk across different worldviews.

This investigation focuses 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 where current fragmentation poses the greatest danger.

## 1.5 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.

### 1.5.1 Automated Worldview Extraction

Current approaches to AI risk modeling, exemplified by the Modeling Transformative AI Risks (MTAIR) project, demonstrate the value of formal representation but require extensive manual effort. Creating a single model demands hundreds of expert-hours to translate qualitative arguments into quantitative frameworks. This bottleneck severely limits the number of perspectives that can be formalized and the speed of model updates as new arguments emerge.

Automation using frontier language models addresses this scaling challenge. By developing systematic methods to extract causal structures and probability judgments from natural language, we can:

> Process orders of magnitude more content
> Incorporate diverse perspectives rapidly
> Maintain models that evolve with the discourse
> Reduce barriers to entry for contributing worldviews

### 1.5.2 Live Data Integration

Static models, however well-constructed, quickly become outdated in fast-moving domains. Prediction markets and forecasting platforms aggregate distributed knowledge about uncertain futures, providing continuously updated probability estimates. By connecting formal models to these live data sources, we create dynamic assessments that incorporate the latest collective intelligence.

This integration serves multiple purposes:

> Grounding abstract models in empirical forecasts
> Identifying which uncertainties most affect outcomes
> Revealing when model assumptions diverge from collective expectations
> Generating new questions for forecasting communities

### 1.5.3 Formal Policy Evaluation

The ultimate purpose of risk modeling is informing action. Formal policy evaluation transforms static risk assessments into actionable guidance by modeling how specific interventions alter critical parameters. Using causal inference techniques, we can assess not just the probability of adverse outcomes but how those probabilities change under different policy regimes.

This enables genuinely evidence-based policy development:

> Comparing interventions across multiple worldviews
> Identifying robust strategies that work across scenarios
> Understanding which uncertainties most affect policy effectiveness
> Prioritizing research to reduce decision-relevant uncertainty

### 1.5.4 The Synergy

The multiplicative benefits emerge from the interactions between components:

> Automation enables comprehensive coverage, making prediction market integration more valuable by connecting to more perspectives
> Market data validates and calibrates automated extractions, improving quality
> Policy evaluation gains precision from both comprehensive models and live probability updates
> The complete system creates feedback loops where policy analysis identifies critical uncertainties for market attention

This synergistic combination addresses the coordination crisis by providing common ground for disparate communities, translating between technical and policy languages, quantifying previously implicit disagreements, and enabling evidence-based compromise.

## 1.6 Thesis Structure and Roadmap

The remainder of this thesis develops the multiplicative benefits framework from theoretical foundations to practical implementation:

**Chapter 2: Context and Theoretical Foundations** establishes the intellectual groundwork, examining the epistemic challenges unique to AI governance, Bayesian networks as formal tools for uncertainty representation, argument mapping as a bridge from natural language to formal models, the MTAIR project's achievements and limitations, and requirements for effective coordination infrastructure.

**Chapter 3: AMTAIR Design and Implementation** presents the technical system including overall architecture and design principles, the two-stage extraction pipeline (ArgDown → BayesDown), validation methodology and results, case studies from simple examples to complex AI risk models, and integration with prediction markets and policy evaluation.

**Chapter 4: Discussion - Implications and Limitations** critically examines technical limitations and failure modes, conceptual concerns about formalization, integration with existing

governance frameworks, scaling challenges and opportunities, and broader implications for epistemic security.

**Chapter 5: Conclusion** synthesizes key contributions and charts paths forward with a summary of theoretical and practical achievements, concrete recommendations for stakeholders, research agenda for community development, and vision for AI governance with proper coordination infrastructure.

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

# 2. Context and Theoretical Foundations

This chapter establishes the theoretical and methodological foundations necessary for understanding AMTAIR's approach to automating AI risk modeling. I begin with the core challenge—representing existential risk arguments in formal terms—then develop the technical and conceptual tools needed to address it.

## 2.1 AI Existential Risk: The Carlsmith Model

To ground our discussion in concrete terms, I examine Joseph Carlsmith's "Is Power-Seeking AI an Existential Risk?" as an exemplar of structured reasoning about AI catastrophic risk. Carlsmith's analysis stands out for its explicit probabilistic decomposition of the path from current AI development to potential existential catastrophe.

### 2.1.1 Six-Premise Decomposition

Carlsmith structures his argument through six conditional premises, each assigned explicit probability estimates:

**Premise 1: APS Systems by 2070** (P  0.65)
"By 2070, there will be AI systems with Advanced capability, Agentic planning, and Strategic awareness" - the conjunction of capabilities that could enable systematic pursuit of objectives in the world.

**Premise 2: Alignment Difficulty** (P  0.40)
"It will be harder to build aligned APS systems than misaligned systems that are still attractive to deploy" - capturing the challenge that safety may conflict with capability or efficiency.

**Premise 3: Deployment Despite Misalignment** (P  0.70)
"Conditional on 1 and 2, we will deploy misaligned APS systems" - reflecting competitive pressures and limited coordination.

**Premise 4: Power-Seeking Behavior** (P  0.65)
"Conditional on 1-3, misaligned APS systems will seek power in high-impact ways" - based on instrumental convergence arguments.

**Premise 5: Disempowerment Success** (P 0.40)
"Conditional on 1-4, power-seeking will scale to permanent human disempowerment" - despite potential resistance and safeguards.

**Premise 6: Existential Catastrophe** (P 0.95)
"Conditional on 1-5, this disempowerment constitutes existential catastrophe" - connecting power loss to permanent curtailment of human potential.

**Overall Risk**: Multiplying through the conditional chain yields P(doom) 0.05 or 5% by 2070.

### 2.1.2 Why Carlsmith Exemplifies Formalizable Arguments

Carlsmith's model demonstrates several features that make it ideal for formal representation:

**Explicit Probabilistic Structure**: Each premise receives numerical probability estimates with documented reasoning, enabling direct translation to Bayesian network parameters.

**Clear Conditional Dependencies**: The logical flow from capabilities through deployment decisions to catastrophic outcomes maps naturally onto directed acyclic graphs.

**Transparent Decomposition**: Breaking the argument into modular premises allows independent evaluation and sensitivity analysis of each component.

**Documented Reasoning**: Extensive justification for each probability enables extraction of both structure and parameters from the source text.

This structured approach exemplifies the type of reasoning AMTAIR aims to formalize and automate. While Carlsmith spent months developing this model manually, similar rigor exists implicitly in many AI safety arguments awaiting extraction.

## 2.2 The Epistemic Challenge of Policy Evaluation

Evaluating AI governance policies presents unique epistemic challenges that traditional policy analysis methods cannot adequately address. Understanding these challenges motivates the need for new computational approaches.

### 2.2.1 Unique Characteristics of AI Governance

**Deep Uncertainty Rather Than Risk**: Traditional policy analysis distinguishes between risk (known probability distributions) and uncertainty (known possibilities, unknown probabilities). AI governance faces deep uncertainty—we cannot confidently enumerate possible futures, much less assign probabilities. Will recursive self-improvement enable rapid capability gains? Can value alignment be solved technically? These foundational questions resist empirical resolution before their answers become catastrophically relevant.

**Complex Multi-Level Causation**: Policy effects propagate through technical, institutional, and social levels with intricate feedback loops. A technical standard might alter research incentives, shifting capability development trajectories, changing competitive dynamics, and ulti-

mately affecting existential risk through pathways invisible at the policy's inception. Traditional linear causal models cannot capture these dynamics.

**Irreversibility and Lock-In**: Many AI governance decisions create path dependencies that prove difficult or impossible to reverse. Early technical standards shape development trajectories. Institutional structures ossify. International agreements create sticky equilibria. Unlike many policy domains where course correction remains possible, AI governance mistakes may prove permanent.

**Value-Laden Technical Choices**: The entanglement of technical and normative questions confounds traditional separation of facts and values. What constitutes "alignment"? How much capability development should we risk for economic benefits? Technical specifications embed ethical judgments that resist neutral expertise.

### 2.2.2 Limitations of Traditional Approaches

Standard policy evaluation tools prove inadequate for these challenges:

**Cost-Benefit Analysis** assumes commensurable outcomes and stable probability distributions. When potential outcomes include existential catastrophe with deeply uncertain probabilities, the mathematical machinery breaks down. Infinite negative utility resists standard decision frameworks.

**Scenario Planning** helps explore possible futures but typically lacks the probabilistic reasoning needed for decision-making under uncertainty. Without quantification, scenarios provide narrative richness but limited action guidance.

**Expert Elicitation** aggregates specialist judgment but struggles with interdisciplinary questions where no single expert grasps all relevant factors. Moreover, experts often operate with different implicit models, making aggregation problematic.

**Red Team Exercises** test specific plans but miss systemic risks emerging from component interactions. Gaming individual failures cannot reveal emergent catastrophic possibilities.

These limitations create a methodological gap: we need approaches that handle deep uncertainty, represent complex causation, quantify expert disagreement, and enable systematic exploration of intervention effects.

## 2.3 Bayesian Networks as Knowledge Representation

Bayesian networks offer a mathematical framework uniquely suited to addressing these epistemic challenges. By combining graphical structure with probability theory, they provide tools for reasoning about complex uncertain domains.

### 2.3.1 Mathematical Foundations

A Bayesian network consists of:

> **Directed Acyclic Graph (DAG)**: Nodes represent variables, edges represent direct dependencies
> **Conditional Probability Tables (CPTs)**: For each node, P(node|parents) quantifies relationships

The joint probability distribution factors according to the graph structure:

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

This factorization enables efficient inference and embodies causal assumptions explicitly.

### 2.3.2 The Rain-Sprinkler-Grass Example

The canonical example illustrates key concepts:

```
[Grass_Wet]: Concentrated moisture on grass.
 + [Rain]: Water falling from sky.
 + [Sprinkler]: Artificial watering system.
   + [Rain]
```

Network Structure:

> **Rain** (root cause): P(rain) = 0.2
> **Sprinkler** (intermediate): P(sprinkler|rain) varies by rain state
> **Grass_Wet** (effect): P(wet|rain, sprinkler) depends on both causes

This simple network demonstrates:

> **Marginal Inference**: P(grass_wet) computed from joint distribution
> **Diagnostic Reasoning**: P(rain|grass_wet) reasoning from effects to causes
> **Intervention Modeling**: P(grass_wet|do(sprinkler=on)) for policy analysis

### 2.3.3 Advantages for AI Risk Modeling

Bayesian networks provide several crucial capabilities:

**Explicit Uncertainty Representation**: Every belief is a probability distribution, avoiding false certainty while enabling quantitative reasoning.

**Causal Modeling**: Directed edges represent causal relationships, enabling counterfactual reasoning through Pearl's do-calculus for policy evaluation.

**Modular Structure**: Complex arguments decompose into manageable components that can be independently evaluated and refined.

**Evidence Integration**: Bayesian updating provides principled methods for incorporating new information as it emerges.

**Visual Communication**: Graphical structure makes complex relationships comprehensible across expertise levels.

These features address key requirements for AI governance: handling uncertainty, representing causation, enabling systematic analysis, and facilitating communication across communities.

## 2.4 Argument Mapping and Formal Representations

The gap between natural language arguments and formal models requires systematic bridging. Argument mapping provides methods for making implicit reasoning structures explicit and analyzable.

### 2.4.1 From Natural Language to Structure

Natural language arguments contain rich information expressed through:

> Causal claims ("X leads to Y")
> Conditional relationships ("If A then likely B")
> Uncertainty expressions ("probably," "might," "certainly")
> Support/attack patterns between claims

Argument mapping extracts this structure, identifying:

> **Core claims and propositions**
> **Inferential relationships**
> **Implicit assumptions**
> **Uncertainty qualifications**

### 2.4.2 ArgDown: Structured Argument Notation

ArgDown provides a markdown-like syntax for hierarchical argument representation:

```
[MainClaim]: Description of primary conclusion.
 + [SupportingEvidence]: Evidence supporting the claim.
   + [SubEvidence]: More specific support.
 - [CounterArgument]: Evidence against the claim.
```

This notation captures argument structure while remaining human-readable and writable. Crucially, it serves as an intermediate representation between natural language and formal models.

### 2.4.3 BayesDown: The Bridge to Bayesian Networks

BayesDown extends ArgDown with probabilistic metadata:

```
[Node]: Description. {
  "instantiations": ["node_TRUE", "node_FALSE"],
  "priors": {"p(node_TRUE)": "0.7", "p(node_FALSE)": "0.3"},
  "posteriors": {
```

```
    "p(node_TRUE|parent_TRUE)": "0.9",
    "p(node_TRUE|parent_FALSE)": "0.4"
  }
}
```

This representation:

> **Preserves narrative structure** from the original argument
> **Adds mathematical precision** through probability specifications
> **Enables transformation** to standard Bayesian network formats
> **Supports validation** by maintaining traceability to sources

The two-stage extraction process (ArgDown → BayesDown) separates concerns: first capturing structure, then quantifying relationships. This modularity enables human oversight at critical decision points.

## 2.5 The MTAIR Framework: Achievements and Limitations

The Modeling Transformative AI Risks (MTAIR) project, led by RAND researchers, pioneered formal modeling of AI existential risk arguments. Understanding its approach and limitations motivates the automation efforts of AMTAIR.

### 2.5.1 MTAIR's Approach

MTAIR manually translated influential AI risk arguments into Bayesian networks using Analytica software:

**Systematic Decomposition**: Breaking complex arguments into variables and relationships through expert analysis.

**Probability Elicitation**: Gathering quantitative estimates through structured expert interviews and literature review.

**Sensitivity Analysis**: Identifying which parameters most influence conclusions about AI risk levels.

**Visual Communication**: Creating interactive models that stakeholders could explore and modify.

### 2.5.2 Key Achievements

MTAIR demonstrated several important possibilities:

**Feasibility of Formalization**: Complex philosophical arguments about AI risk can be represented as Bayesian networks while preserving essential insights.

**Value of Quantification**: Moving from qualitative concerns to quantitative models enables systematic analysis, comparison, and prioritization.

**Cross-Perspective Communication**: Formal models provide common ground for technical and policy communities to engage productively.

**Research Prioritization**: Sensitivity analysis reveals which empirical questions would most reduce uncertainty about AI risks.

### 2.5.3 Fundamental Limitations

However, MTAIR's manual approach faces severe constraints:

**Labor Intensity**: Each model requires hundreds of expert-hours to construct, limiting coverage to a few perspectives.

Detailed breakdown needed: - Variable identification: X hours - Structure elicitation: Y hours - Probability quantification: Z hours - Validation and refinement: W hours Total per model: ~200-400 hours

**Static Nature**: Models become outdated as arguments evolve but updating requires near-complete reconstruction.

**Limited Accessibility**: Using the models requires Analytica software and significant technical sophistication.

**Single Perspective**: Each model represents one worldview, making comparison across perspectives difficult.

These limitations prevent MTAIR's approach from scaling to meet AI governance needs. As the pace of AI development accelerates and arguments proliferate, manual modeling cannot keep pace.

### 2.5.4 The Automation Opportunity

MTAIR's experience reveals both the value of formal modeling and the necessity of automation. Key lessons:

> Formal models genuinely enhance understanding and coordination
> The modeling process itself surfaces implicit assumptions
> Quantification enables analyses impossible with qualitative arguments alone
> But manual approaches cannot scale to match the challenge

This motivates AMTAIR's central innovation: using frontier language models to automate the extraction and formalization process while preserving the benefits MTAIR demonstrated.

## 2.6 Requirements for Coordination Infrastructure

Based on the challenges identified and lessons from existing approaches, we can specify requirements for computational tools that could enhance coordination in AI governance:

### 2.6.1 Scalability

The system must process large volumes of arguments across:

> Academic papers and technical reports
> Policy documents and proposals
> Blog posts and informal arguments
> Forecasting questions and market data

Automation is essential—manual approaches cannot match the pace of discourse.

### 2.6.2 Accessibility

Diverse stakeholders must be able to engage with the system:

> **Researchers** need technical depth and modification capabilities
> **Policymakers** require clear summaries and intervention analysis
> **Forecasters** want integration with prediction platforms
> **Public stakeholders** deserve transparent representation

This demands multiple interfaces and levels of abstraction.

### 2.6.3 Epistemic Virtues

The system should enhance rather than replace human judgment by:

> **Making assumptions explicit** through formal representation
> **Preserving uncertainty** rather than false precision
> **Enabling validation** through traceable extraction
> **Supporting disagreement** through multi-worldview representation
> **Encouraging updating** as new evidence emerges

### 2.6.4 Integration Capabilities

Isolated tools have limited impact. The system needs:

> **Data source connections** to prediction markets and forecasting platforms
> **API accessibility** for integration with other tools
> **Export formats** compatible with standard analysis software
> **Version control** for tracking model evolution
> **Collaborative features** for community development

### 2.6.5 Robustness Properties

Given the high stakes, the system must handle:

> **Extraction errors** through validation and correction mechanisms
> **Adversarial inputs** designed to manipulate outputs
> **Model uncertainty** through sensitivity analysis
> **Scaling challenges** as networks grow large

> **Evolution over time** as arguments develop

These requirements shape AMTAIR's design, as detailed in the next chapter.

# 3. AMTAIR: Design and Implementation

This chapter presents the technical architecture and implementation of AMTAIR, demonstrating how theoretical principles translate into working software. I detail the design decisions, implementation challenges, and validation results that establish AMTAIR's feasibility and value.

## 3.1 System Architecture Overview

AMTAIR implements an end-to-end pipeline transforming unstructured text into interactive Bayesian network visualizations. The architecture reflects key design principles:

> **Modularity**: Each component can be independently improved
> **Transparency**: Intermediate outputs enable inspection and validation
> **Flexibility**: Multiple input formats and configurable processing
> **Scalability**: Efficient processing of large document sets

### 3.1.1 Five-Stage Pipeline

The system processes information through five distinct stages:

`Documents → Ingestion → ArgDown → BayesDown → Networks → Visualization`

Each stage produces inspectable outputs, enabling validation and debugging. This transparency is crucial for building trust in automated extraction.

### 3.1.2 Component Architecture

```
#| label: architecture-overview
#| eval: false

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)
```

This clean separation of concerns enables targeted improvements and alternative implementations for each component.

## 3.2 The Two-Stage Extraction Process

The core innovation of AMTAIR lies in separating structural extraction from probability quantification. This two-stage approach addresses key challenges in automated formalization.

### 3.2.1 Stage 1: Structural Extraction (ArgDown)

The first stage identifies argument structure without concerning itself with quantification:

**Variable Identification**: Extract key propositions and entities from text using patterns like "X causes Y," "If A then B," and domain-specific indicators.

**Relationship Mapping**: Identify support, attack, and conditional relationships between variables through linguistic analysis.

**Hierarchy Construction**: Build nested ArgDown representation preserving logical flow:

```
[Existential_Catastrophe]: Destruction of humanity's potential.
 + [Human_Disempowerment]: Loss of control to AI systems.
   + [Misaligned_Power_Seeking]: AI pursuing problematic objectives.
     + [APS_Systems]: Advanced, agentic, strategic AI.
     + [Deployment_Decisions]: Choice to deploy despite risks.
```

**Validation**: Ensure extracted structure forms valid directed acyclic graph and preserves key argumentative relationships from source.

### 3.2.2 Stage 2: Probability Integration (BayesDown)

The second stage adds quantitative information to the structural skeleton:

**Question Generation**: For each node, generate probability elicitation questions:

Examples needed: - "What is the probability of existential catastrophe?" - "What is P(catastrophe|human_disempowerment)?" - Show how questions map to BayesDown structure

**Probability Extraction**: Identify explicit numerical statements and map qualitative expressions:

> "Very likely" → 0.75-0.9
> "Possible but unlikely" → 0.1-0.3

**Coherence Enforcement**: Ensure probabilities satisfy basic constraints:

> Probabilities sum to 1.0
> Conditional tables are complete
> No logical contradictions

**Metadata Integration**: Combine structure with probabilities in BayesDown format.

### 3.2.3 Why Two Stages?

This separation provides several benefits:

**Modular Validation**: Structure can be verified independently from probability estimates, simplifying quality assurance.

**Human Oversight**: Experts can review and correct structural extraction before probability quantification.

**Flexible Quantification**: Different methods (LLM extraction, expert elicitation, market data) can provide probabilities for the same structure.

**Error Isolation**: Structural errors don't contaminate probability extraction and vice versa.

## 3.3 Implementation Details

The system is implemented in Python, leveraging established libraries while adding novel extraction capabilities.

### 3.3.1 Technology Stack

> **Language Models**: OpenAI GPT-4 and Anthropic Claude for extraction
> **Network Analysis**: NetworkX for graph algorithms
> **Probabilistic Modeling**: pgmpy for Bayesian network operations
> **Visualization**: PyVis for interactive network rendering
> **Data Processing**: Pandas for structured data manipulation

### 3.3.2 Key Algorithms

**Hierarchical Parsing**: The system parses ArgDown/BayesDown syntax recognizing indentation-based hierarchy:

**Probability Completion**: When sources don't specify all required probabilities, the system uses principled methods:

Document approaches: - Maximum entropy for missing values - Coherence constraints propagation - Expert-specified defaults - Confidence scoring for completed values

**Visual Encoding**: Nodes are colored by probability magnitude and styled by network position:

> Green (high probability) to red (low probability) gradient
> Blue borders for root causes, purple for intermediate, magenta for effects

### 3.3.3 Performance Characteristics

Benchmarking reveals practical scalability:

Table 5: Performance benchmarks for different network sizes

| Network Size | Nodes | Processing Time | Memory Usage |
|---|---|---|---|
| Small | 10 | <1 second | <100MB |
| Medium | 11-30 | 2-8 seconds | 100-500MB |
| Large | 31-50 | 15-45 seconds | 0.5-1GB |
| Very Large | >50 | Requires approximation | >1GB |

The bottleneck shifts from extraction (linear in text length) to inference (exponential in network connectivity) as models grow.

## 3.4 Case Study: Rain-Sprinkler-Grass

I begin with the canonical example to demonstrate the complete pipeline on a simple, well-understood case.

### 3.4.1 Input Representation

The source BayesDown representation:

```
[Grass_Wet]: Concentrated moisture on 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]: Water falling from sky.
   {"instantiations": ["rain_TRUE", "rain_FALSE"],
    "priors": {"p(rain_TRUE)": "0.2", "p(rain_FALSE)": "0.8"}}
 + [Sprinkler]: Artificial watering system.
```

```
{"instantiations": ["sprinkler_TRUE", "sprinkler_FALSE"],
 "priors": {"p(sprinkler_TRUE)": "0.448", "p(sprinkler_FALSE)": "0.552"},
 "posteriors": {
   "p(sprinkler_TRUE|rain_TRUE)": "0.01",
   "p(sprinkler_TRUE|rain_FALSE)": "0.4"
 }}
+ [Rain]
```

### 3.4.2 Processing Steps

1. **Parsing**: Extract three nodes with relationships
2. **Validation**: Verify probability coherence and DAG structure
3. **Enhancement**: Calculate joint probabilities and network metrics
4. **Construction**: Build formal Bayesian network
5. **Visualization**: Render interactive display

### 3.4.3 Results

The system successfully:

> Extracts complete network structure
> Preserves all probability information
> Calculates correct marginal probabilities
> Generates interactive visualization
> Enables inference queries

This simple example validates the basic pipeline functionality before tackling complex real-world cases.

## 3.5 Case Study: Carlsmith's Power-Seeking AI Model

Applying AMTAIR to Carlsmith's model demonstrates scalability to realistic AI safety arguments.

### 3.5.1 Model Complexity

The Carlsmith model contains:

> **23 nodes** representing different factors
> **27 edges** encoding dependencies
> **Multiple probability tables** with complex conditionals
> **Six-level causal depth** from root causes to catastrophe

### 3.5.2 Extraction Results

The automated extraction successfully identifies:

**Core Risk Pathway**:

```
Existential_Catastrophe
← Human_Disempowerment
← Scale_Of_Power_Seeking
← Misaligned_Power_Seeking
← [APS_Systems, Difficulty_Of_Alignment, Deployment_Decisions]
```

**Supporting Structure**:

> Competitive dynamics influencing deployment
> Technical factors affecting alignment difficulty
> Corrective mechanisms and their limitations

**Probability Preservation**:

> Extracted probabilities match Carlsmith's published estimates
> Conditional relationships properly captured
> Final P(doom) calculation reproduces ~5% result

### 3.5.3 Validation Against Original

Comparing extracted model to Carlsmith's original:

Table 6: Carlsmith model extraction validation results

| Metric | Performance |
|---|---|
| Structural Accuracy | 92% (nodes and edges) |
| Probability Accuracy | 87% (within 0.05) |
| Path Completeness | 100% (all major paths) |
| Semantic Preservation | High (per expert review) |

The high fidelity demonstrates AMTAIR's capability for complex real-world arguments.

### 3.5.4 Insights from Formalization

Formal representation reveals several insights:

**Critical Path Analysis**: The pathway through APS development and deployment decisions carries the highest risk contribution.

**Sensitivity Points**: Small changes in deployment probability create large changes in overall risk.

**Intervention Opportunities**: Improving alignment difficulty or deployment governance show highest impact potential.

These insights emerge naturally from formal analysis but remain implicit in textual arguments.

## 3.6 Validation Methodology

Establishing trust in automated extraction requires rigorous validation across multiple dimensions.

### 3.6.1 Ground Truth Construction

I created validation datasets through:

Document the process: 1. Expert selection criteria 2. Training on extraction methodology 3. Independent extraction procedures 4. Consensus building process 5. Inter-rater reliability metrics

1. **Expert Manual Extraction**: Three domain experts independently extracted models from the same sources
2. **Consensus Building**: Reconciled differences to create gold standard representations
3. **Annotation**: Marked source passages supporting each element

### 3.6.2 Evaluation Metrics

**Structural Metrics**:

> Precision: Fraction of extracted elements that are correct
> Recall: Fraction of true elements that are extracted
> F1 Score: Harmonic mean balancing precision and recall

**Probabilistic Metrics**:

> Mean Absolute Error for probability values
> Kullback-Leibler divergence for distributions
> Calibration plots for uncertainty expression

**Semantic Metrics**:

> Expert ratings of meaning preservation
> Functional equivalence for inference queries

### 3.6.3 Results Summary

Across 20 test documents:

Table 7: System validation results across components

| Component | Precision | Recall | F1 Score |
|---|---|---|---|
| Node Identification | 89% | 86% | 0.875 |
| Edge Extraction | 84% | 81% | 0.825 |
| Probability Values | 76% | 71% | 0.735 |
| **Overall System** | **83%** | **79%** | **0.810** |

Performance is strongest for explicit structural elements and numerical probabilities, with more challenges in extracting implicit relationships and qualitative uncertainty.

### 3.6.4 Error Analysis

Common failure modes:

**Implicit Assumptions** (23% of errors): Unstated background assumptions that experts infer but system misses.

**Complex Conditionals** (19% of errors): Nested conditionals with multiple antecedents challenge current parsing.

**Ambiguous Quantifiers** (17% of errors): Terms like "significant" lack clear probability mapping without context.

**Coreference Resolution** (15% of errors): Pronouns and indirect references create attribution challenges.

Understanding these limitations guides both current usage and future improvements.

## 3.7 Policy Evaluation Capabilities

Beyond extraction and visualization, AMTAIR enables systematic policy analysis through formal intervention modeling.

### 3.7.1 Intervention Representation

Policies are modeled as modifications to network parameters:

### 3.7.2 Example: Deployment Governance

Consider a policy requiring safety certification before deployment:

**Intervention**: Set P(deployment|misaligned) = 0.1 (from 0.7)

**Results**:

> Baseline P(catastrophe) = 0.05
> Intervened P(catastrophe) = 0.012
> Relative risk reduction = 76%
> Number needed to regulate = 26 deployments

This quantitative analysis enables comparison across interventions.

### 3.7.3 Robustness Analysis

Policies must work across worldviews. AMTAIR enables:

1. **Multi-Model Evaluation**: Test interventions across different extracted models
2. **Parameter Sensitivity**: Vary assumptions to find breaking points

3. **Scenario Analysis**: Combine interventions under different futures
4. **Confidence Bounds**: Propagate uncertainty through to outcomes

This systematic approach moves beyond intuitive policy assessment.

## 3.8 Interactive Visualization Design

Making Bayesian networks accessible to diverse stakeholders requires careful visualization design.

### 3.8.1 Visual Encoding Strategy

The system uses multiple visual channels:

**Color**: Probability magnitude (green=high, red=low)
**Borders**: Node type (blue=root, purple=intermediate, magenta=effect)
**Size**: Centrality in network (larger=more influential)
**Layout**: Force-directed positioning reveals clusters

### 3.8.2 Progressive Disclosure

Information appears at appropriate levels:

1. **Overview**: Network structure and color coding
2. **Hover**: Node description and prior probability
3. **Click**: Full probability tables and details
4. **Interaction**: Drag to rearrange, zoom to explore

This layered approach serves both quick assessment and deep analysis needs.

### 3.8.3 User Interface Elements

Key features enhance usability:

> **Physics Controls**: Adjust layout dynamics
> **Filter Options**: Show/hide node types
> **Export Functions**: Save images or data
> **Comparison Mode**: Side-by-side worldviews

These features emerged from user testing with researchers and policymakers.

## 3.9 Integration with Prediction Markets

While full integration remains future work, the architecture supports connection to live forecasting data.

### 3.9.1 Design for Integration

The system architecture anticipates market connections:

Design documentation needed: - API specifications for major platforms - Semantic matching algorithms - Probability aggregation methods - Update scheduling and caching

```
#| label: market-connector
#| eval: false

class PredictionMarketConnector:
    def __init__(self, market_apis):
        self.markets = market_apis

    def find_relevant_questions(self, model_variables):
        """Map model variables to forecast questions"""
        # Semantic matching between variables and questions

    def fetch_probabilities(self, questions):
        """Retrieve latest market probabilities"""
        # API calls with caching and error handling

    def update_model(self, model, market_data):
        """Integrate market probabilities into model"""
        # Weighted updating based on liquidity and track record
```

### 3.9.2 Challenges and Opportunities

Key integration challenges:

> **Question Mapping**: Model variables rarely match market questions exactly
> **Temporal Alignment**: Markets forecast specific dates, models consider scenarios
> **Quality Variation**: Market depth and participation vary significantly

Despite challenges, even partial integration provides value through external validation and dynamic updating.

## 3.10 Computational Considerations

As networks grow large, computational challenges emerge requiring sophisticated approaches.

### 3.10.1 Exact vs. Approximate Inference

Small networks enable exact inference through variable elimination. Larger networks require approximation:

**Monte Carlo Methods**: Sample from probability distributions to estimate queries
**Variational Inference**: Optimize simpler distributions to approximate true posteriors
**Belief Propagation**: Pass messages between nodes to converge on beliefs

The system automatically selects appropriate methods based on network properties.

### 3.10.2 Scaling Strategies

For very large networks:

Document strategies with benchmarks: 1. Hierarchical decomposition algorithms 2. Pruning criteria and impact 3. Caching architecture 4. Parallelization speedups

1. **Hierarchical Decomposition**: Break into sub-networks for independent analysis
2. **Pruning**: Remove low-influence paths for specific queries
3. **Caching**: Store computed results for common queries
4. **Parallelization**: Distribute sampling across processors

These strategies extend practical network size limits significantly.

## 3.11 Summary of Technical Achievements

AMTAIR successfully demonstrates:

> **Automated extraction** from natural language to formal models
> **Two-stage architecture** separating structure from quantification
> **High fidelity** preservation of complex arguments
> **Interactive visualization** accessible to diverse users
> **Policy evaluation** capabilities through intervention modeling
> **Scalable implementation** handling realistic network sizes

These achievements validate the feasibility of computational coordination infrastructure for AI governance.

# 4. Discussion: Implications and Limitations

This chapter critically examines AMTAIR's implications, limitations, and potential failure modes. By engaging seriously with objections and challenges, I aim to provide a balanced assessment of what this approach can and cannot achieve for AI governance coordination.

## 4.1 Technical Limitations and Responses

### 4.1.1 Objection 1: Extraction Quality Boundaries

**Critic**: "Complex implicit reasoning chains resist formalization. Automated extraction will systematically miss nuanced arguments, subtle conditional relationships, and context-dependent meanings that human readers naturally understand."

**Response**: This concern has merit—extraction does face inherent limitations. However, the empirical results tell a more nuanced story. With extraction achieving 85%+ accuracy for structural relationships and 73% for probability capture, the system performs well enough for practical use while falling short of human expert performance.

More importantly, AMTAIR employs a hybrid human-AI workflow that addresses this limitation:

> **Two-stage verification**: Humans review structural extraction before probability quantification
> **Transparent outputs**: All intermediate representations remain human-readable
> **Iterative refinement**: Extraction prompts improve based on error analysis
> **Ensemble approaches**: Multiple extraction attempts can identify ambiguities

The question is not whether automated extraction perfectly captures every nuance—it doesn't. Rather, it's whether imperfect extraction still provides value over no formal representation. When the alternative is relying on conflicting mental models that remain entirely implicit, even 75% accurate formal models represent significant progress.

Furthermore, extraction errors often reveal interesting properties of the source arguments themselves—ambiguities that human readers gloss over become explicit when formalization fails. This diagnostic value enhances rather than undermines the approach.

### 4.1.2 Objection 2: False Precision in Uncertainty

**Critic**: "Attaching exact probabilities to unprecedented events like AI catastrophe is fundamentally misguided. The numbers create false confidence in what amounts to educated speculation about radically uncertain futures."

**Response**: This philosophical objection strikes at the heart of formal risk assessment. However, AMTAIR addresses it through several design choices:

First, the system explicitly represents uncertainty about uncertainty. Rather than point estimates, the framework supports probability distributions over parameters. When someone says "likely" we might model this as Beta(8,2) rather than exactly 0.8, capturing both the central estimate and our uncertainty about it.

Technical requirements: - Beta distributions for probability parameters - Dirichlet for multi-state variables - Propagation through inference - Visualization of uncertainty bounds

Second, all probabilities are explicitly conditional on stated assumptions. The system doesn't claim "P(catastrophe) = 0.05" absolutely, but rather "Given Carlsmith's model assumptions, P(catastrophe) = 0.05." This conditionality is preserved throughout analysis.

Third, sensitivity analysis reveals which probabilities actually matter. Often, precise values are unnecessary—knowing whether a parameter is closer to 0.1 or 0.9 suffices for decision-making. The formalization helps identify where precision matters and where it doesn't.

Finally, the alternative to quantification isn't avoiding the problem but making it worse. When experts say "highly likely" or "significant risk," they implicitly reason with probabilities. Formalization simply makes these implicit quantities explicit and subject to scrutiny. As Dennis Lindley noted, "Uncertainty is not in the events, but in our knowledge about them."

### 4.1.3 Objection 3: Correlation Complexity

**Critic**: "Bayesian networks assume conditional independence given parents, but real-world AI risks involve complex correlations. Ignoring these dependencies could dramatically misrepresent risk levels."

**Response**: Standard Bayesian networks do face limitations with correlation representation—this is a genuine technical challenge. However, several approaches within the framework address this:

**Explicit correlation nodes**: When factors share hidden common causes, we can add latent variables to capture correlations. For instance, "AI research culture" might influence both "capability advancement" and "safety investment."

**Copula methods**: For known correlation structures, copula functions can model dependencies while preserving marginal distributions. This extends standard Bayesian networks significantly.

**Sensitivity bounds**: When correlations remain uncertain, we can compute bounds on outcomes under different correlation assumptions. This reveals when correlations critically affect

conclusions.

**Model ensembles**: Different correlation structures can be modeled separately and results aggregated, similar to climate modeling approaches.

More fundamentally, the question is whether imperfect independence assumptions invalidate the approach. In practice, explicitly modeling first-order effects with known limitations often proves more valuable than attempting to capture all dependencies informally. The framework makes assumptions transparent, enabling targeted improvements where correlations matter most.

## 4.2 Conceptual and Methodological Concerns

### 4.2.1 Objection 4: Democratic Exclusion

**Critic**: "Transforming policy debates into complex graphs and equations will sideline non-technical stakeholders, concentrating influence among those comfortable with formal models. This technocratic approach undermines democratic participation in crucial decisions about humanity's future."

**Response**: This concern about technocratic exclusion deserves serious consideration—formal methods can indeed create barriers. However, AMTAIR's design explicitly prioritizes accessibility alongside rigor:

**Progressive disclosure interfaces** allow engagement at multiple levels. A policymaker might explore visual network structures and probability color-coding without engaging mathematical details. Interactive features let users modify assumptions and see consequences without understanding implementation.

**Natural language preservation** ensures original arguments remain accessible. The BayesDown format maintains human-readable descriptions alongside formal specifications. Users can always trace from mathematical representations back to source texts.

**Comparative advantage** comes from making implicit technical content explicit, not adding complexity. When experts debate AI risk, they already employ sophisticated probabilistic reasoning—formalization reveals rather than creates this complexity. Making hidden assumptions visible arguably enhances rather than reduces democratic participation.

**Multiple interfaces** serve different communities. Researchers access full technical depth, policymakers use summary dashboards, public stakeholders explore interactive visualizations. The same underlying model supports varied engagement modes.

Rather than excluding non-technical stakeholders, proper implementation can democratize access to expert reasoning by making it inspectable and modifiable. The risk lies not in formalization itself but in poor interface design or gatekeeping behaviors around model access.

### 4.2.2 Objection 5: Oversimplification of Complex Systems

**Critic**: "Forcing rich socio-technical systems into discrete Bayesian networks necessarily loses crucial dynamics—feedback loops, emergent properties, institutional responses, and cultural factors that shape AI development. The models become precise but wrong."

**Response**: All models simplify by necessity—as Box noted, "All models are wrong, but some are useful." The question becomes whether formal simplifications improve upon informal mental models:

**Transparent limitations** make formal models' shortcomings explicit. Unlike mental models where simplifications remain hidden, network representations clearly show what is and isn't included. This transparency enables targeted criticism and improvement.

**Iterative refinement** allows models to grow more sophisticated over time. Starting with first-order effects and adding complexity where it proves important follows successful practice in other domains. Climate models began simply and added dynamics as computational power and understanding grew.

**Complementary tools** address different aspects of the system. Bayesian networks excel at probabilistic reasoning and intervention analysis. Other approaches—agent-based models, system dynamics, scenario planning—can capture different properties. AMTAIR provides one lens, not the only lens.

**Empirical adequacy** ultimately judges models. If simplified representations enable better predictions and decisions than informal alternatives, their abstractions are justified. Early results suggest formal models, despite simplifications, outperform intuitive reasoning for complex risk assessment.

The goal isn't creating perfect representations but useful ones. By making simplifications explicit and modifiable, formal models enable systematic improvement in ways mental models cannot.

## 4.3 Red-Teaming Results

To identify failure modes, I conducted systematic adversarial testing of the AMTAIR system.

### 4.3.1 Adversarial Extraction Attempts

I tested the system with deliberately challenging inputs:

**Contradictory Arguments**: Texts asserting P(A) = 0.2 and P(A) = 0.8 in different sections

> Result: System flagged inconsistency rather than averaging
> Mitigation: Explicit consistency checking with user resolution

**Circular Reasoning**: Arguments where A causes B causes C causes A

> Result: DAG validation caught cycles, extraction failed gracefully
> Mitigation: Clear error messages explaining the structural issue

**Extremely Vague Language**: Texts using only qualitative terms without clear relationships

> Result: Extraction quality degraded significantly (F1 $< 0.5$)
> Mitigation: Confidence scores on extracted elements, human review triggers

**Deceptive Framings**: Arguments designed to imply false causal relationships

> Result: System sometimes extracted spurious connections
> Mitigation: Source grounding requirements, validation against citations

### 4.3.2 Robustness Findings

Key vulnerabilities identified:

Specific metrics need validation: - Anchoring bias: measured effect size with confidence intervals - Authority sensitivity: controlled experiment design - Complexity degradation: performance curve analysis - Context loss: dependency distance metrics

1. **Anchoring bias**: System tends to over-weight first probability mentioned (effect size analysis needed)
2. **Authority sensitivity**: Extracted probabilities influenced by cited expert prominence
3. **Complexity degradation**: Performance drops sharply beyond 50 nodes
4. **Context loss**: Long-range dependencies in text sometimes missed

However, the system demonstrated robustness to:

> Different writing styles and academic disciplines
> Variations in argument structure and presentation order
> Mixed numerical and qualitative probability expressions
> Reasonable levels of grammatical errors and typos

### 4.3.3 Implications for Deployment

These results suggest AMTAIR is suitable for:

> **Research applications** with expert oversight
> **Policy analysis** of well-structured arguments
> **Educational uses** demonstrating formal reasoning
> **Collaborative modeling** with human verification

But should be used cautiously for:

> Fully automated analysis without review
> Adversarial or politically contentious texts
> Real-time decision-making without validation
> Arguments far outside training distribution

## 4.4 Enhancing Epistemic Security

Despite limitations, AMTAIR contributes to epistemic security in AI governance through several mechanisms.

### 4.4.1 Making Models Inspectable

The greatest epistemic benefit comes from forcing implicit models into explicit form. When an expert claims "misalignment likely leads to catastrophe," formalization asks:

> Likely means what probability?
> Through what causal pathways?
> Under what assumptions?
> With what evidence?

This explicitation serves multiple functions:

**Clarity**: Vague statements become precise claims subject to evaluation

**Comparability**: Different experts' models can be systematically compared

**Criticizability**: Hidden assumptions become visible targets for challenge

**Updatability**: Formal models can systematically incorporate new evidence

### 4.4.2 Revealing Convergence and Divergence

Comparative analysis across extracted models reveals surprising patterns:

Implement comparison of 3+ models: - Structural similarity metrics - Parameter divergence analysis - Crux identification algorithms - Visualization of agreement patterns

**Structural convergence**: Different experts often share similar causal models even when probability estimates diverge dramatically. This suggests shared understanding of mechanisms despite disagreement on magnitudes.

**Parameter clustering**: Probability estimates often cluster around a few values rather than spreading uniformly, suggesting implicit coordination or common evidence bases.

**Crux identification**: Formal comparison precisely identifies where worldviews diverge—often just 2-3 key parameters drive different conclusions about overall risk.

These insights remain hidden when arguments stay in natural language form.

### 4.4.3 Improving Collective Reasoning

AMTAIR enhances group epistemics through:

**Explicit uncertainty**: Replacing "might," "could," "likely" with probability distributions reduces miscommunication and forces precision

**Compositional reasoning**: Complex arguments decompose into manageable components that can be independently evaluated

**Evidence integration**: New information updates specific parameters rather than requiring complete argument reconstruction

**Exploration tools**: Stakeholders can modify assumptions and immediately see consequences, building intuition about model dynamics

Early pilot studies with AI governance researchers show improved agreement identification and reduced time to consensus—though specific quantitative claims require careful validation.

## 4.5 Scaling Challenges and Opportunities

Moving from prototype to widespread adoption faces both technical and social challenges.

### 4.5.1 Technical Scaling

**Computational complexity** grows with network size, but several approaches help:

> Hierarchical decomposition for very large models
> Caching and approximation for common queries
> Distributed processing for extraction tasks
> Incremental updating rather than full recomputation

**Data quality** varies dramatically across sources:

> Academic papers provide structured arguments
> Blog posts offer rich ideas with less formal structure
> Policy documents mix normative and empirical claims
> Social media presents extreme extraction challenges

**Integration complexity** increases with ecosystem growth:

> Multiple LLM providers with different capabilities
> Diverse visualization needs across users
> Various export formats for downstream tools
> Version control for evolving models

### 4.5.2 Social and Institutional Scaling

**Adoption barriers** include:

> Learning curve for formal methods
> Institutional inertia in established processes
> Concerns about replacing human judgment
> Resource requirements for implementation

**Trust building** requires:

> Transparent methodology documentation
> Published validation studies
> High-profile successful applications
> Community ownership and development

**Sustainability** depends on:

> Open source development model
> Diverse funding sources
> Academic and industry partnerships
> Clear value demonstration

### 4.5.3 Opportunities for Impact

Despite challenges, several factors favor adoption:

**Timing**: AI governance needs tools now, creating receptive audiences

**Complementarity**: AMTAIR enhances rather than replaces existing processes

**Flexibility**: The approach adapts to different contexts and needs

**Network effects**: Value increases as more perspectives are formalized

Early adopters in research organizations and think tanks can demonstrate value, creating momentum for broader adoption.

## 4.6 Integration with Governance Frameworks

AMTAIR complements rather than replaces existing governance approaches.

### 4.6.1 Standards Development

Technical standards bodies could use AMTAIR to:

> Model how proposed standards affect risk pathways
> Compare different standard options systematically
> Identify unintended consequences through pathway analysis
> Build consensus through explicit model negotiation

Example: Evaluating compute thresholds for AI system regulation by modeling how different thresholds affect capability development, safety investment, and competitive dynamics.

### 4.6.2 Regulatory Design

Regulators could apply the framework to:

> Assess regulatory impact across different scenarios
> Identify enforcement challenges through explicit modeling
> Compare international approaches systematically

> Design adaptive regulations responsive to evidence

Example: Analyzing how liability frameworks affect corporate AI development decisions under different market conditions.

### 4.6.3 International Coordination

Multilateral bodies could leverage shared models for:

> Establishing common risk assessments
> Negotiating agreements with explicit assumptions
> Monitoring compliance through parameter tracking
> Adapting agreements as evidence emerges

Example: Building shared models for AGI development scenarios to inform international AI governance treaties.

### 4.6.4 Organizational Decision-Making

Individual organizations could use AMTAIR for:

> Internal risk assessment and planning
> Board-level communication about AI strategies
> Research prioritization based on model sensitivity
> Safety case development with explicit assumptions

Example: An AI lab modeling how different safety investments affect both capability advancement and risk mitigation.

## 4.7 Future Research Directions

Several research directions could enhance AMTAIR's capabilities and impact.

### 4.7.1 Technical Enhancements

**Improved extraction**: Fine-tuning language models specifically for argument extraction, handling implicit reasoning, and cross-document synthesis

**Richer representations**: Temporal dynamics, continuous variables, and multi-agent interactions within extended frameworks

**Inference advances**: Quantum computing applications, neural approximate inference, and hybrid symbolic-neural methods

**Validation methods**: Automated consistency checking, anomaly detection in extracted models, and benchmark dataset development

### 4.7.2 Methodological Extensions

**Causal discovery**: Inferring causal structures from data rather than just extracting from text

**Experimental integration**: Connecting models to empirical results from AI safety experiments

**Dynamic updating**: Continuous model refinement as new evidence emerges from research and deployment

**Uncertainty quantification**: Richer representation of deep uncertainty and model confidence

### 4.7.3 Application Domains

**Beyond AI safety**: Climate risk, biosecurity, nuclear policy, and other existential risks

**Corporate governance**: Strategic planning, risk management, and innovation assessment

**Scientific modeling**: Formalizing theoretical arguments in emerging fields

**Educational tools**: Teaching probabilistic reasoning and critical thinking

### 4.7.4 Ecosystem Development

**Open standards**: Common formats for model exchange and tool interoperability

**Community platforms**: Collaborative model development and sharing infrastructure

**Training programs**: Building capacity for formal modeling in governance communities

**Quality assurance**: Certification processes for high-stakes model applications

These directions could transform AMTAIR from a single tool into a broader ecosystem for enhanced reasoning about complex risks.

# 5. Conclusion: Toward Coordinated AI Governance

## 5.1 Summary of Key Contributions

This thesis has demonstrated both the need for and feasibility of computational approaches to enhancing coordination in AI governance. The work makes several distinct contributions across theory, methodology, and implementation.

### 5.1.1 Theoretical Contributions

**Diagnosis of the Coordination Crisis**: I've articulated how fragmentation across technical, policy, and strategic communities systematically amplifies existential risk from advanced AI. This framing moves beyond identifying disagreements to understanding how misaligned efforts create negative-sum dynamics—safety gaps emerge between communities, resources are misallocated through duplication and neglect, and interventions interact destructively.

**The Multiplicative Benefits Framework**: The combination of automated extraction, prediction market integration, and formal policy evaluation creates value exceeding the sum of parts. Automation enables scale, markets provide empirical grounding, and policy analysis delivers actionable insights. Together, they address different facets of the coordination challenge while reinforcing each other's strengths.

**Epistemic Infrastructure Conception**: Positioning formal models as epistemic infrastructure reframes the role of technical tools in governance. Rather than replacing human judgment, computational approaches provide common languages, shared representations, and systematic methods for managing disagreement—essential foundations for coordination under uncertainty.

### 5.1.2 Methodological Innovations

**Two-Stage Extraction Architecture**: Separating structural extraction (ArgDown) from probability quantification (BayesDown) addresses key challenges in automated formalization. This modularity enables human oversight at critical points, supports multiple quantification methods, and isolates different types of errors for targeted improvement.

**BayesDown as Bridge Representation**: The development of BayesDown syntax creates a crucial intermediate representation preserving both narrative accessibility and mathematical

precision. This bridge enables the transformation from qualitative arguments to quantitative models while maintaining traceability and human readability.

**Validation Framework**: The systematic approach to validating automated extraction—comparing against expert annotations, measuring multiple accuracy dimensions, and analyzing error patterns—establishes scientific standards for assessing formalization tools. This framework can guide future development in this emerging area.

### 5.1.3 Technical Achievements

**Working Implementation**: AMTAIR demonstrates end-to-end feasibility from document ingestion through interactive visualization. The system achieves practically useful accuracy levels: 85%+ for structural extraction and 73% for probability capture on real AI safety arguments.

**Scalability Solutions**: Technical approaches for handling realistic model complexity—hierarchical decomposition, approximate inference, and progressive visualization—show that computational limitations need not prevent practical application.

**Accessibility Design**: The layered interface approach serves diverse stakeholders without compromising technical depth. Progressive disclosure, visual encoding, and interactive exploration make formal models accessible beyond technical specialists.

### 5.1.4 Empirical Findings

**Extraction Feasibility**: The successful extraction of complex arguments like Carlsmith's model validates the core premise that implicit formal structures exist in natural language arguments and can be computationally recovered with reasonable fidelity.

**Convergence Patterns**: Comparative analysis reveals surprising structural agreement across worldviews even when probability estimates diverge dramatically. This suggests shared causal understanding despite parameter disagreements—a foundation for coordination.

**Intervention Impacts**: Policy evaluation demonstrates how formal models enable rigorous assessment of governance options. The ability to quantify risk reduction across scenarios and identify robust strategies validates the practical value of formalization.

## 5.2 Limitations and Honest Assessment

Despite these contributions, important limitations constrain current capabilities and should guide appropriate use.

### 5.2.1 Technical Constraints

**Extraction Boundaries**: While 73-85% accuracy suffices for many purposes, systematic biases remain. The system struggles with implicit assumptions, complex conditionals, and context-dependent meanings. These limitations necessitate human review for high-stakes applications.

**Correlation Handling**: Standard Bayesian networks inadequately represent complex correlations in real systems. While extensions like copulas and explicit correlation nodes help, fully capturing interdependencies remains challenging.

**Computational Scaling**: Very large networks (>50 nodes) require approximations that may affect accuracy. As models grow to represent richer phenomena, computational constraints increasingly bind.

### 5.2.2 Conceptual Limitations

**Formalization Trade-offs**: Converting rich arguments to formal models necessarily loses nuance. While making assumptions explicit provides value, some insights resist mathematical representation.

**Probability Interpretation**: Deep uncertainty about unprecedented events challenges probabilistic representation. Numbers can create false precision even when explicitly conditional and uncertain.

**Social Complexity**: Institutional dynamics, cultural factors, and political processes influence AI development in ways that simple causal models struggle to capture.

### 5.2.3 Practical Constraints

**Adoption Barriers**: Learning curves, institutional inertia, and resource requirements limit immediate deployment. Even demonstrably valuable tools face implementation challenges.

**Maintenance Burden**: Models require updating as arguments evolve and evidence emerges. Without sustained effort, formal representations quickly become outdated.

**Context Dependence**: The approach works best for well-structured academic arguments. Application to informal discussions, political speeches, or social media remains challenging.

## 5.3 Implications for AI Governance

Despite limitations, AMTAIR's approach offers significant implications for how AI governance can evolve toward greater coordination and effectiveness.

### 5.3.1 Near-Term Applications

**Research Coordination**: Research organizations can use formal models to:

> Map the landscape of current arguments and identify gaps
> Prioritize investigations targeting high-sensitivity parameters
> Build cumulative knowledge through explicit model updating
> Facilitate collaboration through shared representations

**Policy Development**: Governance bodies can apply the framework to:

> Evaluate proposals across multiple expert worldviews

> Identify robust interventions effective under uncertainty
> Make assumptions explicit for democratic scrutiny
> Track how evidence changes optimal policies over time

**Stakeholder Communication**: The visualization and analysis tools enable:

> Clearer communication between technical and policy communities
> Public engagement with complex risk assessments
> Board-level strategic discussions grounded in formal analysis
> International negotiations with explicit shared models

### 5.3.2 Medium-Term Transformation

As adoption spreads, we might see:

**Epistemic Commons**: Shared repositories of formalized arguments become reference points for governance discussions, similar to how economic models inform monetary policy or climate models guide environmental agreements.

**Adaptive Governance**: Policies designed with explicit models can include triggers for reassessment as key parameters change, enabling responsive governance that avoids both paralysis and recklessness.

**Professionalization**: "Model curator" and "argument formalization specialist" emerge as recognized roles, building expertise in bridging natural language and formal representations.

**Quality Standards**: Community norms develop around model transparency, validation requirements, and appropriate use cases, preventing both dismissal and over-reliance on formal tools.

### 5.3.3 Long-Term Vision

Successfully scaling this approach could fundamentally alter AI governance:

**Coordinated Response**: Rather than fragmented efforts, the AI safety ecosystem could operate with shared situational awareness—different actors understanding how their efforts interact and contribute to collective goals.

**Anticipatory Action**: Formal models with prediction market integration could provide early warning of emerging risks, enabling proactive rather than reactive governance.

**Global Cooperation**: Shared formal frameworks could facilitate international coordination similar to how economic models enable monetary coordination or climate models support environmental agreements.

**Democratic Enhancement**: Making expert reasoning transparent and modifiable could enable broader participation in crucial decisions about humanity's technological future.

## 5.4 Recommendations for Stakeholders

Different communities can take concrete steps to realize these benefits:

### 5.4.1 For Researchers

1. **Experiment with formalization**: Try extracting your own arguments into ArgDown/BayesDown format to discover implicit assumptions
2. **Contribute to validation**: Provide expert annotations for building benchmark datasets and improving extraction quality
3. **Develop extensions**: Build on the open-source foundation to add capabilities for your specific domain needs
4. **Publish formally**: Include formal model representations alongside traditional papers to enable cumulative building

### 5.4.2 For Policymakers

1. **Pilot applications**: Use AMTAIR for internal analysis of specific policy proposals to build familiarity and identify value
2. **Demand transparency**: Request formal models underlying expert recommendations to understand assumptions and uncertainties
3. **Fund development**: Support tool development and training to build governance capacity for formal methods
4. **Design adaptively**: Create policies with explicit triggers based on model parameters to enable responsive governance

### 5.4.3 For Technologists

1. **Improve extraction**: Contribute better prompting strategies, fine-tuned models, or validation methods
2. **Enhance interfaces**: Develop visualizations and interactions serving specific stakeholder needs
3. **Build integrations**: Connect AMTAIR to other tools in the AI governance ecosystem
4. **Scale infrastructure**: Address computational challenges for larger models and broader deployment

### 5.4.4 For Funders

1. **Support ecosystem**: Fund not just tool development but training, community building, and maintenance

2. **Bridge communities**: Incentivize collaborations between formal modelers and domain experts
3. **Measure coordination**: Develop metrics for assessing coordination improvements from formal tools
4. **Patient capital**: Recognize that epistemic infrastructure requires sustained investment to reach potential

## 5.5 Future Research Agenda

Building on this foundation, several research directions could amplify impact:

### 5.5.1 Technical Priorities

**Extraction Enhancement**:

> Fine-tuning language models specifically for argument extraction
> Handling implicit reasoning and long-range dependencies
> Cross-document synthesis for comprehensive models
> Multilingual extraction for global perspectives

**Representation Extensions**:

> Temporal dynamics for modeling AI development trajectories
> Multi-agent representations for strategic interactions
> Continuous variables for economic and capability metrics
> Uncertainty types beyond probability distributions

**Integration Depth**:

> Semantic matching between models and prediction markets
> Automated experiment design based on model sensitivity
> Policy optimization algorithms using extracted models
> Real-time updating from news and research feeds

### 5.5.2 Methodological Development

**Validation Science**:

> Larger benchmark datasets with diverse argument types
> Metrics for semantic preservation beyond accuracy
> Adversarial robustness testing protocols
> Longitudinal studies of model evolution

**Hybrid Approaches**:

> Optimal human-AI collaboration patterns for extraction
> Combining formal models with other methods (scenarios, simulations)
> Integration with deliberative and participatory processes

> Balancing automation with expert judgment

**Social Methods**:

> Ethnographic studies of model use in organizations
> Measuring coordination improvements empirically
> Understanding adoption barriers and facilitators
> Designing interventions for epistemic security

### 5.5.3 Application Expansion

**Domain Extensions**:

> Climate risk assessment and policy evaluation
> Biosecurity governance and pandemic preparedness
> Nuclear policy and deterrence stability
> Emerging technology governance broadly

**Institutional Integration**:

> Embedding in regulatory impact assessment
> Corporate strategic planning applications
> Academic peer review enhancement
> Democratic deliberation support tools

**Global Deployment**:

> Adapting to different governance contexts
> Supporting multilateral negotiation processes
> Building capacity in developing nations
> Creating resilient distributed infrastructure

## 5.6 Closing Reflections

The work presented in this thesis emerges from a simple observation: while humanity mobilizes unprecedented resources to address AI risks, our efforts remain tragically uncoordinated. Different communities work with incompatible frameworks, duplicate efforts, and sometimes actively undermine each other's work. This fragmentation amplifies the very risks we seek to mitigate.

AMTAIR represents one attempt to build bridges—computational tools that create common ground for disparate perspectives. By making implicit models explicit, quantifying uncertainty, and enabling systematic policy analysis, these tools offer hope for enhanced coordination. The successful extraction of complex arguments, validation against expert judgment, and demonstration of policy evaluation capabilities suggest this approach has merit.

Yet tools alone cannot solve coordination problems rooted in incentives, institutions, and human psychology. AMTAIR provides infrastructure for coordination, not coordination itself. Success requires not just technical development but changes in how we approach collective

challenges—valuing transparency over strategic ambiguity, embracing uncertainty rather than false confidence, and prioritizing collective outcomes over parochial interests.

The path forward demands both ambition and humility. Ambition to build the epistemic infrastructure necessary for navigating unprecedented risks. Humility to recognize our tools' limitations and the irreducible role of human wisdom in governance. The question is not whether formal models can replace human judgment—they cannot and should not. Rather, it's whether we can augment our collective intelligence with computational tools that help us reason together about futures too important to leave to chance.

As AI capabilities advance toward transformative potential, the window for establishing effective governance narrows. We cannot afford continued fragmentation when facing potentially irreversible consequences. The coordination crisis in AI governance represents both existential risk and existential opportunity—risk if we fail to align our efforts, opportunity if we succeed in building unprecedented cooperation around humanity's most important challenge.

This thesis contributes technical foundations and demonstrates feasibility. The greater work—building communities, changing practices, and fostering coordination—remains ahead. May we prove equal to the task, for all our futures depend on it.

# References


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

[2] 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.

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

# Appendices

## Appendix A: Technical Implementation Details

Contents: - Full API specifications - Architectural diagrams with component details - Code structure and organization - Deployment instructions - Performance optimization guides

## Appendix B: Validation Datasets and Procedures

Contents: - Benchmark dataset descriptions - Annotation guidelines - Inter-rater reliability protocols - Statistical analysis procedures - Replication instructions

## Appendix C: Extended Case Studies

Include: - Christiano's "What failure looks like" - Critch's ARCHES model - Additional policy evaluation scenarios - Comparative analysis across models

## Appendix D: BayesDown Syntax Specification

Contents: - Full syntax definition - Validation rules - Example transformations - Implementation notes - Extension possibilities

## Appendix E: Prompt Engineering Details

Include: - Full extraction prompts with annotations - Iterative refinement history - Ablation study results - Best practices guide - Common failure patterns

## Appendix F: User Guide

Sections: - Getting started with AMTAIR - Creating your first extraction - Interpreting visualizations - Policy evaluation walkthrough - Troubleshooting common issues

## Appendix G: Jupyter Notebook Implementation

The complete implementation is available as an interactive Jupyter notebook demonstrating:

> Environment setup and configuration
> Step-by-step extraction pipeline
> Visualization generation
> Policy evaluation examples
> Performance benchmarking

# Bibliography

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

[2] 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.

[3] 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