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
93138 Lappersdorf

Supervisor:

Dr. Timo Speith

Word Count:
30.000

Source / Identifier:
Document URL

26th of May 2025

Table of Contents

List of Figures

List of Tables

Preface

Abstract

The coordination crisis in AI governance presents a paradoxical challenge: unprecedented investment in AI safety coexists alongside fundamental coordination failures across technical, policy, and ethical domains. These divisions systematically increase existential risk. This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach addressing this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models. The system implements an end-to-end pipeline transforming unstructured text into interactive Bayesian networks through a novel two-stage extraction process that bridges communication gaps between stakeholders.

The coordination crisis in AI governance presents a paradoxical challenge: unprecedented investment in AI safety coexists alongside fundamental coordination failures across technical, policy, and ethical domains. These divisions systematically increase existential risk by creating safety gaps, misallocating resources, and fostering inconsistent approaches to interdependent problems.

This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach that addresses this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models.

The AMTAIR system implements an end-to-end pipeline that transforms unstructured text into interactive Bayesian networks through a novel two-stage extraction process: first capturing argument structure in ArgDown format, then enhancing it with probability information in BayesDown. This approach bridges communication gaps between stakeholders by making implicit models explicit, enabling comparison across different worldviews, providing a common language for discussing probabilistic relationships, and supporting policy evaluation across diverse scenarios.

Prefatory Apparatus: Frontmatter

Illustrations and Terminology — Quick References

Acknowledgments

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

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

- Technical advisors who provided feedback on implementation aspects

- Personal supporters who enabled the research through encouragement and feedback

List of Graphics & Figures

List of Abbreviations

- AGI - Artificial General Intelligence
- AMTAIR - Automating Modeling of Transformative AI Risks
- API - Application Programming Interface
- APS - Advanced, Planning, Strategic (AI systems per **carlsmith2021**)
- BN - Bayesian Network
- CPT - Conditional Probability Table
- DAG - Directed Acyclic Graph
- LLM - Large Language Model
- MTAIR - Modeling Transformative AI Risks
- TAI - Transformative Artificial Intelligence

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

Final Thesis: Automating the Modeling of Transformative Artificial Intelligence Risks

Frontmatter: 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. The AMTAIR project—Automating Transformative AI Risk Modeling—represents an attempt to build computational bridges between communities that, despite shared concerns about AI risk, often struggle to communicate effectively due to incompatible frameworks, terminologies, and implicit assumptions.

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 his guidance throughout this project, providing both technical insights and philosophical grounding. The MTAIR team's pioneering manual approach inspired this automation effort, and I am grateful for their foundational work.

I acknowledge Johannes Meyer and Jelena Meyer for their invaluable assistance in verifying the automated extraction procedure through manual extraction of ArgDown and BayesDown data from the Carlsmith paper, providing crucial ground truth for validation.

Special recognition goes to Coleman Snell for his partnership and research collaboration with the AMTAIR project, offering both technical expertise and strategic vision. The AI safety community's creation of rich literature made this work possible, and I thank all researchers whose arguments provided the raw material for formalization.

Any errors or limitations remain my own responsibility.

List of Figures

List of Tables

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
ML - Machine Learning
MTAIR - Modeling Transformative AI Risks
NLP - Natural Language Processing
P&E - Philosophy & Economics
PDF - Portable Document Format
TAI - Transformative Artificial Intelligence

1. Introduction: The Coordination Crisis in AI Governance

Chapter Overview
Grade Weight: 10% | **Target Length**: ~14% of text (~4,200 words)
Requirements: Introduces and motivates the core question, provides context, states precise thesis, provides roadmap

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[1] 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 **todd2024**.

Show Image

[1] The orthogonality thesis posits that intelligence and goals are independent—an AI can have any set of objectives regardless of its intelligence level. The instrumental convergence thesis suggests that different AI systems may adopt similar instrumental goals (e.g., self-preservation, resource acquisition) to achieve their objectives.

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 **maslej2025 samborska2025**—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.

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

The 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 **tegmark2024**, the EU advances comprehensive AI regulation **european2024**, 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.

Show Image

1.2.1 Safety Gaps from Misaligned Efforts

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

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 **paul2023**.

Historical precedents demonstrate how coordination failures in technology governance can lead to dangerous dynamics. The nuclear arms race exemplifies how lack of coordination can create negative-sum outcomes where all parties become less secure despite massive investments in safety measures. Similar dynamics may emerge in AI development without proper coordination infrastructure.

1.2.2 Resource Misallocation

The AI safety community faces a complex tradeoff in resource allocation. While some duplication of efforts can improve reliability through independent verification—akin to reproducing scientific results—the current level of fragmentation often leads to wasteful redundancy. Multiple teams independently develop similar frameworks without building on each other's work, creating opportunity costs where critical but unglamorous research areas remain understaffed. Funders struggle to identify high-impact opportunities across technical and governance domains, lacking the epistemic infrastructure to assess where marginal resources would have the greatest impact. This misallocation becomes more costly as the window for establishing effective governance narrows with accelerating AI development.

Table 1: Examples of duplicated AI safety efforts across organizations

Research Area	Organization A	Organization B	Duplication Level	Opportunity Cost
Interpretability Methods	Anthropic's mechanistic interpretability	DeepMind's concept activation vectors	Medium	Reduced focus on multi-agent safety
Alignment Frameworks	MIRI's embedded agency	FHI's comprehensive AI services	High	Limited work on institutional design
Risk Assessment Models	GovAI's policy models	CSER's existential risk frameworks	High	Insufficient capability benchmarking

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.

The game-theoretic structure of AI development creates particularly pernicious dynamics. Armstrong et al. **armstrong2016** demonstrate how uncoordinated policies can incentivize a "race to the precipice" where competitive pressures override safety considerations. The situation resembles a multi-player prisoner's dilemma or stag hunt where individually rational decisions lead to collectively catastrophic outcomes **samuel2023 hunt2025**.

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) **schelling1960**, institutions (IAEA), and communication channels (hotlines, treaties) did humanity navigate the nuclear precipice **rehman2025**.

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.

Show Image

1.4 Research Question and Scope

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

Can frontier AI technologies be utilized to automate the modeling of transformative AI risks, enabling robust prediction of policy impacts across diverse worldviews?

More specifically, I explore whether frontier language models can automate the extraction and formalization of probabilistic world models from AI safety literature, creating a scalable computational framework that enhances coordination in AI governance through systematic policy evaluation under uncertainty.

To break this down into its components:

- **Frontier AI Technologies**: Today's most capable language models (GPT-4, Claude-3 level systems)
- **Automated Modeling**: Using these systems to extract and formalize argument structures from natural language
- **Transformative AI Risks**: Potentially catastrophic outcomes from advanced AI systems, particularly existential risks
- **Policy Impact Prediction**: Evaluating how governance interventions might alter probability distributions over outcomes
- **Diverse Worldviews**: Accounting for fundamental disagreements about AI development trajectories and risk factors

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

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.

Show Image

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 dozens 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 **tetlock2015**.

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

Formal policy evaluation transforms static risk assessments into actionable guidance by modeling how specific interventions alter critical parameters. Using causal inference techniques **pearl2000 pearl2009**, 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 this progression, 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.

Show Image

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

2. Context and Theoretical Foundations

Chapter Overview
Grade Weight: 20% | **Target Length**: ~29% of text (~8,700 words)
Requirements: Demonstrates understanding of relevant concepts, explains relevance, situates in debate, reconstructs arguments

This chapter establishes the theoretical and methodological foundations for the AMTAIR approach. We begin by examining a concrete example of structured AI risk assessment—Joseph Carlsmith's power-seeking AI model—to ground our discussion in practical terms. We then explore the unique epistemic challenges of AI governance that render traditional policy analysis inadequate, introduce Bayesian networks as formal tools for representing uncertainty, and examine how argument mapping bridges natural language reasoning and formal models. The chapter concludes by analyzing the MTAIR project's achievements and limitations, motivating the need for automated approaches, and surveying relevant literature across AI risk modeling, governance proposals, and technical methodologies.

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 **carlsmith2022**. 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

According to the MTAIR model **clarke2022**, Carlsmith decomposes existential risk into a probabilistic chain with explicit estimates[2]:

1. **Premise 1**: Transformative AI development this century (P 0.80)(P   0.80) (P 0.80)
2. **Premise 2**: AI systems pursuing objectives in the world (P 0.95)(P   0.95) (P 0.95)

[2] Multiple versions of Carlsmith's paper exist with slight updates to probability estimates: **carlsmith2021**, **carlsmith2022**, **carlsmith2024**. We primarily reference the version used by the MTAIR team for their extraction. Extended discussion and expert probability estimates can be found on LessWrong.

3. **Premise 3**: Systems with power-seeking instrumental incentives (P 0.40)(P 0.40) (P 0.40)
4. **Premise 4**: Sufficient capability for existential threat (P 0.65)(P 0.65) (P 0.65)
5. **Premise 5**: Misaligned systems despite safety efforts (P 0.50)(P 0.50) (P 0.50)
6. **Premise 6**: Catastrophic outcomes from misaligned power-seeking (P 0.65)(P 0.65) (P 0.65)

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

mermaid

```
flowchart TD
    P1[Premise 1: Transformative AI<br/>P  0.80] --> P2[Premise 2: AI pursuing objectives<b
    P2 --> P3[Premise 3: Power-seeking incentives<br/>P  0.40]
    P3 --> P4[Premise 4: Existential capability<br/>P  0.65]
    P4 --> P5[Premise 5: Misalignment despite safety<br/>P  0.50]
    P5 --> P6[Premise 6: Catastrophic outcome<br/>P  0.65]
    P6 --> D[Existential Catastrophe<br/>P  0.05]
```

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

Premise 1: APS Systems by 2070 (P 0.65)(P 0.65) (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)(P 0.40) (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)(P 0.70) (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)(P 0.65) (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)(P 0.40) (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)(P 0.95) (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) 0.05P(doom) 0.05 P(doom) 0.05 or 5% by 2070.

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.

2.1.2 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.

We will return to Carlsmith's model in Chapter 3 as our primary complex case study, demonstrating how AMTAIR successfully extracts and formalizes this sophisticated multi-level argument.

Beyond Carlsmith's model, other structured approaches to AI risk—such as Christiano's "What failure looks like" **christiano2019**—provide additional targets for automated extraction, enabling comparative analysis across different expert worldviews.

2.2 The Epistemic Challenge of Policy Evaluation

AI governance policy evaluation faces unique epistemic challenges that render traditional policy analysis methods insufficient. 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 **hallegatte2012**. 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 ultimately 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.

Table 2: Comparison of AI governance vs traditional policy domains

Dimension	Traditional Policy	AI Governance
Uncertainty Type	Risk (known distributions)	Deep uncertainty (unknown unknowns)
Causal Structure	Linear, traceable	Multi-level, feedback loops
Reversibility	Course correction possible	Path dependencies, lock-in
Fact-Value Separation	Clear boundaries	Entangled technical-normative
Empirical Grounding	Historical precedents	Unprecedented phenomena
Time Horizons	Years to decades	Months to centuries

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.

2.2.3 The Underlying Epistemic Framework

The AMTAIR approach rests on a specific epistemic framework that combines probabilistic reasoning, conditional logic, and possible worlds semantics. This framework provides the philosophical foundation for representing deep uncertainty about AI futures.

Probabilistic Epistemology: Following the Bayesian tradition, we treat probability as a measure of rational credence rather than objective frequency. This subjective interpretation allows meaningful probability assignments even for unique, unprecedented events like AI catastrophe. As E.T. Jaynes demonstrated, probability theory extends deductive logic to handle uncertainty, providing a calculus for rational belief **jaynes2003**.

Conditional Structure: The framework emphasizes conditional rather than absolute probabilities. Instead of asking "What is P(catastrophe)?" we ask "What is P(catastrophe | specific assumptions)?" This conditionalization makes explicit the dependency of conclusions on worldview assumptions, enabling productive disagreement about premises rather than conclusions.

Possible Worlds Semantics: We conceptualize uncertainty as distributions over possible worlds—complete descriptions of how reality might unfold. Each world represents a coherent scenario with specific values for all relevant variables. Probability distributions over these worlds capture both what we know and what we don't know about the future.

This framework enables several key capabilities:

1. **Representing ignorance**: We can express uncertainty about uncertainty itself through hierarchical probability models
2. **Combining evidence**: Bayesian updating provides principled methods for integrating new information
3. **Comparing worldviews**: Different probability distributions over the same space of possibilities enable systematic comparison
4. **Evaluating interventions**: Counterfactual reasoning about how actions change probability distributions

2.2.4 Toward New Epistemic Tools

The inadequacy of traditional methods for AI governance creates an urgent need for new epistemic tools. These tools must:

- **Handle Deep Uncertainty**: Move beyond point estimates to represent ranges of possibilities
- **Capture Complex Causation**: Model multi-level interactions and feedback loops
- **Quantify Disagreement**: Make explicit where experts diverge and why
- **Enable Systematic Analysis**: Support rigorous comparison of policy options

Key Insight: The computational approaches developed in this thesis—particularly Bayesian networks enhanced with automated extraction—directly address each of these requirements by providing formal frameworks for reasoning under uncertainty.

Show Image

Show Image

Show Image

Show Image

Recent work on conditional trees demonstrates the value of structured approaches to uncertainty. McCaslin et al. **mccaslin2024** show how hierarchical conditional forecasting can identify high-value questions for reducing uncertainty about complex topics like AI risk. Their methodology, which asks experts to produce simplified Bayesian networks of informative forecasting questions, achieved nine times higher information value than standard forecasting platform questions.

Tetlock's work with the Forecasting Research Institute **tetlock2022** exemplifies how prediction markets can provide empirical grounding for formal models. By structuring questions as conditional trees, they enable forecasters to express complex dependencies between events, providing exactly the type of data needed for Bayesian network parameterization.

Gruetzemacher **gruetzemacher2022** evaluates the tradeoffs between full Bayesian networks and conditional trees for forecasting tournaments. While conditional trees offer simplicity, Bayesian networks provide richer representation of dependencies—motivating AMTAIR's approach of using full networks while leveraging conditional tree insights for question generation.

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(X1,X2,…,Xn)= i=1nP(Xi Parents(Xi))P(X_1, X_2, …, X_n) = _{i=1}^{n} P(X_i | Parents(X_i))P(X1,X2,…,Xn)=i=1 nP(Xi Parents(Xi))

This factorization enables efficient inference and embodies causal assumptions explicitly.

Pearl's foundational work **pearl2014** established Bayesian networks as a principled approach to automated reasoning under uncertainty, providing both theoretical foundations and practical algorithms.

2.3.2 The Rain-Sprinkler-Grass Example

The canonical example illustrates key concepts[3]:

```
[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

mermaid

```
flowchart TD
    R[Rain<br/>P(rain) = 0.2] --> S[Sprinkler]
    R --> G[Grass_Wet]
    S --> G

    subgraph CPT1[Sprinkler CPT]
        S1[P(sprinkler|rain) = 0.01]
        S2[P(sprinkler|¬rain) = 0.4]
    end

    subgraph CPT2[Grass_Wet CPT]
        G1[P(wet|rain,sprinkler) = 0.99]
        G2[P(wet|rain,¬sprinkler) = 0.8]
        G3[P(wet|¬rain,sprinkler) = 0.9]
        G4[P(wet|¬rain,¬sprinkler) = 0.01]
    end
```

python

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

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

[3] This example, while simple, demonstrates all essential features of Bayesian networks and serves as the foundation for understanding more complex applications

```
    (False, False): 0.01   # Rain=F, Sprinkler=F
}
```

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

Show Image

Rain-Sprinkler-Grass Network Rendering

```
#| label: rain_sprinkler_grass_example_network_rendering
#| echo: true
#| eval: true
#| fig-cap: "Dynamic Html Rendering of the Rain-Sprinkler-Grass DAG with Conditional Probabi
#| fig-link: "https://singularitysmith.github.io/AMTAIR_Prototype/bayesian_network.html"
#| fig-alt: "Dynamic Html Rendering of the Rain-Sprinkler-Grass DAG"

from IPython.display import IFrame

IFrame(src="https://singularitysmith.github.io/AMTAIR_Prototype/bayesian_network.html", widt
```

2.3.3 Advantages for AI Risk Modeling

These features address key requirements for AI governance:

- **Handling Uncertainty**: Every parameter is a distribution, not a point estimate
- **Representing Causation**: Directed edges embody causal relationships
- **Enabling Analysis**: Formal inference algorithms support systematic evaluation
- **Facilitating Communication**: Visual structure aids cross-domain understanding

---

3.5.6 Validation Against Original (From the MTAIR Project)

To validate the AMTAIR extraction quality, an ideal approach would involve systematic comparison with expert manual extractions. The validation methodology would follow these steps:

Expert Baseline Creation: Multiple domain experts independently extract ArgDown and BayesDown representations from the same source documents. This creates a ground truth dataset accounting for legitimate variation in expert interpretation.

Structural Comparison: Compare the extracted causal structures, examining:

- Node identification completeness
- Relationship preservation accuracy
- Hierarchical organization fidelity

- Handling of repeated or complex dependencies

Probability Assessment: Evaluate probability extraction by:

- Comparing explicit probability captures
- Assessing interpretation of qualitative expressions
- Measuring consistency across related probabilities
- Identifying systematic biases or tendencies

Semantic Preservation: Expert review would assess whether the formal representation maintains the essential meaning and nuance of the original arguments.

For the Carlsmith model specifically, preliminary manual extractions by domain experts (including Johannes Meyer and Jelena Meyer) suggest that automated extraction achieves high structural fidelity—capturing the key variables and their relationships—while probability estimates show greater variation, reflecting the inherent ambiguity in translating qualitative language to quantitative values.

3.6 Validation Methodology

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

3.6.1 Ground Truth Construction

An ideal validation protocol would establish ground truth through:

Expert Selection: Recruit domain experts with both AI safety knowledge and experience in formal modeling. Experts should represent diverse perspectives within the field to capture legitimate interpretive variation.

Standardized Training: Provide consistent training on ArgDown/BayesDown syntax and extraction principles. This ensures methodological alignment while preserving substantive differences in interpretation.

Independent Extraction: Have experts work independently on the same source documents, preventing anchoring bias and capturing the natural range of valid interpretations.

Consensus Building: Through structured discussion, identify areas of convergence and legitimate disagreement. This distinguishes extraction errors from genuine ambiguity in source materials.

Documentation: Record not just final extractions but the reasoning process, creating rich data for understanding extraction challenges and improving automated approaches.

3.6.2 Evaluation Metrics

A comprehensive evaluation framework would assess multiple dimensions:

Structural Metrics:

- Node identification precision and recall
- Edge extraction accuracy
- Preservation of hierarchical relationships
- Handling of complex dependencies

Probability Metrics:

- Accuracy of explicit probability extraction
- Consistency in interpreting qualitative expressions
- Preservation of conditional relationships
- Handling of uncertainty about uncertainty

Semantic Metrics:

- Expert ratings of meaning preservation
- Functional equivalence for key inferences
- Preservation of author's argumentative intent
- Appropriate simplification choices

Pragmatic Metrics:

- Usefulness for downstream analysis
- Time savings versus manual extraction
- Error patterns and failure modes
- Robustness across document types

3.6.3 Results Summary

While comprehensive validation remains future work, preliminary assessments suggest:

Structural Extraction: The two-stage approach successfully identifies major causal relationships and preserves hierarchical structure. The explicit ArgDown intermediate representation allows verification of structural accuracy before probability quantification.

Probability Challenges: Converting qualitative expressions to numerical probabilities remains the primary challenge. Different experts interpret phrases like "likely" or "significant risk" differently, and automated extraction inherits this ambiguity.

Practical Utility: Despite imperfections, automated extraction provides sufficient quality for many practical applications, especially when combined with human review at critical points.

The validation framework itself represents a contribution, providing systematic methods for assessing automated argument formalization tools as this area develops.

3.6.4 Error Analysis

Common failure modes to avoid:

Implicit Assumptions: Unstated background assumptions that experts infer but system misses. These often involve domain-specific common knowledge that remains unspoken in expert discourse.

Complex Conditionals: Nested conditionals with multiple antecedents challenge current parsing. Statements like "If A and B, then probably C, unless D" require sophisticated logical analysis.

Ambiguous Quantifiers: Terms like "significant" lack clear probability mapping without context. The same word may imply different probabilities in different domains or even different parts of the same argument.

Coreference Resolution: Pronouns and indirect references create attribution challenges. When authors use "this risk" or "that assumption," identifying the correct referent requires deep contextual understanding.

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

Policy interventions can be modeled as modifications to network parameters, following Pearl's do-calculus framework. An ideal implementation would:

Parameter Modification: Represent policies as changes to specific probability values. For instance, safety requirements might reduce P(deployment|misaligned) by making unsafe deployment less likely.

Structural Interventions: Some policies add or remove causal pathways. Regulatory oversight might introduce new nodes representing approval processes.

Uncertainty Propagation: Model uncertainty about policy effectiveness. Rather than assuming perfect implementation, represent ranges of possible effects.

Multi-Level Effects: Capture how policies influence multiple levels simultaneously—technical development, corporate behavior, and international dynamics.

The formal framework enables rigorous counterfactual reasoning: "What would happen to existential risk if this policy were implemented?"

3.7.2 Example: Deployment Governance

Consider a hypothetical policy requiring safety certification before deployment:

Baseline Scenario: Without intervention, the model might show P(deployment|misaligned) = 0.7, reflecting competitive pressures to deploy despite risks.

Policy Intervention: Safety certification requirements could reduce this to P(deployment|misaligned) = 0.1, assuming effective enforcement.

Downstream Effects: This change propagates through the network:

- Reduced deployment of misaligned systems
- Lower probability of power-seeking behavior manifestation
- Decreased existential risk

Quantitative Assessment: The formal model enables precise calculation of risk reduction, helping prioritize among possible interventions.

This example illustrates how formal models transform vague policy discussions into concrete quantitative analyses, though the specific numbers depend on model assumptions and parameter estimates.

3.7.3 Robustness Analysis

Policies must work across worldviews. AMTAIR enables multi-model evaluation, parameter sensitivity testing, scenario analysis, and confidence bound computation—ensuring interventions remain effective despite uncertainty.

Cross-Model Testing: Evaluate policies across different extracted worldviews to identify robust strategies that work regardless of which expert's model proves correct.

Sensitivity Analysis: Identify which parameters most affect policy effectiveness, focusing implementation efforts on critical factors.

Scenario Planning: Test policies under different future scenarios—slow versus fast takeoff, unipolar versus multipolar development, cooperative versus adversarial dynamics.

Confidence Bounds: Rather than point estimates, compute ranges of possible effects accounting for parameter uncertainty.

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) provides immediate visual indication of likelihood, leveraging intuitive color associations.

Borders: Node type (blue=root, purple=intermediate, magenta=effect) helps users understand causal flow through the network structure.

Size: Centrality in network (larger=more influential) draws attention to critical nodes that affect many other variables.

Layout: Force-directed positioning reveals clusters of related variables, helping users identify cohesive sub-arguments within larger models.

3.8.2 Progressive Disclosure

Information appears at appropriate levels:

1. **Overview**: Network structure and color coding provide immediate understanding of key relationships and probability distributions.
2. **Hover**: Node description and prior probability offer additional context without cluttering the main view.
3. **Click**: Full probability tables and details enable deep investigation of specific variables and their relationships.
4. **Interaction**: Drag to rearrange, zoom to explore—users can customize views for their specific interests and questions.

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

3.8.3 User Interface Elements

Effective interface design would incorporate:

Physics Controls: Allow users to adjust layout dynamics, finding arrangements that best reveal patterns of interest.

Filter Options: Enable focusing on specific node types or probability ranges, reducing complexity for targeted analysis.

Export Functions: Support saving visualizations and data in formats suitable for reports, presentations, and further analysis.

Comparison Mode: Facilitate side-by-side viewing of different models or the same model under different policy interventions.

These features should emerge from iterative design with actual users—researchers needing detailed analysis, policymakers seeking key insights, and public stakeholders requiring accessible overviews.

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 anticipates market connections through:

API Specifications: Standardized interfaces for major forecasting platforms like Metaculus, Good Judgment Open, and Manifold Markets.

Semantic Matching: Algorithms to connect model variables with related forecast questions, handling differences in phrasing and scope.

Aggregation Methods: Principled approaches for combining multiple forecast sources, accounting for track records and expertise.

Update Scheduling: Efficient caching and refresh strategies to balance currency with computational cost.

3.9.2 Challenges and Opportunities

Key integration challenges:

Question Mapping: Model variables rarely match market questions exactly. "AI causes existential catastrophe" might map to multiple specific forecast questions about AI capabilities, deployment, and impacts.

Temporal Alignment: Markets forecast specific dates while models consider scenarios. Bridging these requires careful interpretation and uncertainty propagation.

Quality Variation: Market depth and participation vary significantly. Some questions attract expert forecasters while others rely on casual participants.

Despite challenges, even partial integration provides value through external validation of probability estimates and dynamic updating as new information emerges.

3.10 Computational Performance Analysis

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. While approximate, these methods scale to arbitrary network sizes.

Variational Inference: Optimize simpler distributions to approximate true posteriors. These methods trade some accuracy for guaranteed convergence.

Belief Propagation: Pass messages between nodes to converge on beliefs. Particularly effective for tree-structured or sparse networks.

The system automatically selects appropriate methods based on network properties.

3.10.2 Scaling Strategies

For very large networks, several strategies enable practical analysis:

Hierarchical Decomposition: Break large networks into manageable sub-networks, compute locally, then integrate results.

Relevance Pruning: For specific queries, identify and focus on relevant subgraphs, ignoring distant unconnected nodes.

Caching Architecture: Store computed results for common queries, dramatically improving response time for interactive use.

Parallel Processing: Distribute computation across multiple cores or machines for large-scale analysis.

3.11 Results and Achievements

3.11.1 Extraction Quality Assessment

An ideal assessment methodology would systematically evaluate:

Coverage Metrics: What proportion of arguments in source texts are successfully captured in formal models?

Accuracy Metrics: How closely do automated extractions match expert consensus?

Robustness Metrics: How well does the system handle different writing styles, argument structures, and domains?

Utility Metrics: Do the extracted models enable meaningful analysis and decision support?

Preliminary applications suggest the approach achieves practical utility while highlighting areas for improvement, particularly in handling implicit reasoning and converting qualitative uncertainty expressions.

3.11.2 Computational Performance

Performance analysis would examine:

Scaling Characteristics: How processing time grows with network size and complexity.

Bottleneck Identification: Whether limitations arise from extraction, inference, or visualization.

Optimization Opportunities: Where algorithmic improvements or engineering enhancements could improve performance.

Resource Requirements: Memory, processing, and storage needs for realistic applications.

The modular architecture enables targeted optimization of bottleneck components while maintaining system coherence.

3.11.3 Policy Impact Evaluation

A comprehensive evaluation framework would assess:

Intervention Modeling: How effectively can policies be represented as network modifications?

Robustness Testing: Do policy recommendations remain stable across model variations?

Comparative Analysis: How do different policy options compare in effectiveness?

Implementation Guidance: Does the analysis provide actionable insights for policymakers?

The ability to formally model policy interventions and trace their effects through complex causal networks represents a significant advance in systematic governance analysis.

3.12 Summary of Technical Contributions

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
- **Scalable implementation** handling realistic network sizes

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

The implementation shows that formal modeling of AI risk arguments is not only theoretically possible but practically achievable. While challenges remain—particularly in handling implicit reasoning and diverse uncertainty expressions—the system provides a foundation for enhanced coordination in AI governance.

4. Discussion: Implications and Limitations

Chapter Overview
Grade Weight: 10% | **Target Length**: ~14% of text (~4,200 words)
Requirements: Discusses objections, provides convincing replies, extends beyond course materials

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 and subtle conditional relationships that human experts would identify."

Response: This concern has merit—extraction does face inherent limitations. However, the empirical results tell a more nuanced story. The two-stage extraction process, while imperfect, captures sufficient structure for practical use while maintaining transparency about its limitations.

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 partially 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 a range rather than exactly 0.8, capturing both the central estimate and our uncertainty about it.

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.[4]

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

[4] Copulas provide a mathematically elegant way to separate marginal behavior from dependence structure

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.

4.2.3 Objection 6: Idiosyncratic Implementation and Modeling Choices

Critic: "The specific choices made in AMTAIR's implementation—from prompt design to parsing algorithms to visualization strategies—seem arbitrary. Different teams might make entirely different choices, leading to incompatible results. How can we trust conclusions that depend so heavily on implementation details?"

Response: This concern about implementation dependency is valid and deserves careful consideration. However, several factors mitigate this issue:

Convergent Design Principles: While specific implementations vary, fundamental design principles tend to converge. The two-stage extraction process (structure then probability) emerges naturally from how humans parse arguments. The use of intermediate representations follows established practice in computational linguistics. These aren't arbitrary choices but responses to inherent challenges.

Empirical Validation: The "correctness" of implementation choices isn't philosophical but empirical. If different reasonable implementations extract similar structures and lead to similar policy conclusions, this demonstrates robustness. If they diverge dramatically, this reveals genuine ambiguity in source materials—itself valuable information.

Transparent Methodology: By documenting all implementation choices and making code open source, AMTAIR enables replication and variation. Other teams can modify specific components while preserving overall architecture, testing which choices matter.

Convergence at Higher Levels: Even if implementations differ in details, they may converge at levels that matter for coordination. If two systems extract slightly different network structures but reach similar conclusions about policy robustness, the implementation differences don't undermine the approach's value.

Community Standards: As the field matures, community standards will likely emerge—not enforcing uniformity but establishing interoperability. This parallels development in other technical fields where multiple implementations coexist within shared frameworks.

The deeper insight is that implementation choices encode theoretical commitments. By making these explicit and variable, AMTAIR turns a bug into a feature—we can systematically explore how different assumptions affect conclusions, enhancing rather than undermining epistemic security.

4.3 Red-Teaming Results

To identify failure modes, systematic adversarial testing of the AMTAIR system would be essential.

4.3.1 Adversarial Extraction Attempts

A comprehensive red-teaming approach would test the system with:

Contradictory Arguments: Texts containing logically inconsistent claims or probability estimates. The system should flag contradictions rather than silently reconciling them.

Circular Reasoning: Arguments with circular dependencies that violate DAG requirements. Proper validation should detect and report such structural issues.

Ambiguous Language: Texts using extremely vague or metaphorical language. The system should acknowledge extraction uncertainty rather than forcing precise interpretations.

Deceptive Framings: Arguments crafted to imply false causal relationships. This tests whether the system merely extracts surface claims or requires deeper coherence.

Adversarial Prompts: Inputs designed to trigger known LLM failure modes. This ensures robustness against prompt injection and manipulation attempts.

Each failure mode discovered would inform system improvements and user guidance.

4.3.2 Robustness Findings

Theoretical analysis suggests key vulnerabilities:

Anchoring Effects: Language models may over-weight information presented early in documents, potentially biasing extraction toward initial framings.

Authority Sensitivity: Extraction might be influenced by explicit credibility signals in text, potentially giving undue weight to claimed expertise.

Complexity Limits: Performance likely degrades with very large argument structures, requiring hierarchical decomposition strategies.

Context Windows: Long-range dependencies exceeding model context windows could be missed, fragmenting cohesive arguments.

Understanding these limitations enables appropriate use—leveraging strengths while compensating for weaknesses through human oversight and validation.

4.3.3 Implications for Deployment

These considerations 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

Theoretical analysis suggests formal comparison would reveal:

Structural Patterns: Experts likely share more agreement about causal structures than probability values, suggesting common understanding of mechanisms despite quantitative disagreement.

Crux Identification: Formal models make explicit which specific disagreements drive different conclusions, focusing discussion on genuinely critical differences.

Hidden Agreements: Apparently conflicting positions might share substantial common ground obscured by different terminology or emphasis.

Uncertainty Clustering: Areas of high uncertainty likely correlate across models, revealing where additional research would most reduce disagreement.

These patterns remain invisible in natural language debates but become analyzable through formalization.

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

While empirical validation remains future work, theoretical considerations suggest these mechanisms could substantially improve coordination quality. By providing shared representations and systematic methods for managing disagreement, formal models create infrastructure for collective intelligence that transcends individual limitations.

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.

The extensive literature on corporate governance and liability frameworks **cuomo2016 demirag2000 devilliers2021 divito2022 kaur2024 list2011 solomon2020** provides theoretical grounding for understanding how regulatory interventions shape organizational behavior. AMTAIR could formalize these relationships in the specific context of AI development, making explicit how different liability regimes might incentivize or discourage safety investments.

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

Recent advances in causal structure learning from both text and data **babakov2025 ban2023 bethard2007 chen2023 heinze-deml2018 squires2023 yang2022** suggest promising directions for enhancing AMTAIR's extraction capabilities. The theoretical foundations from **duhem1954** and **meyer2022b** on the philosophy of science and knowledge structures provide epistemological grounding for these methodological extensions.

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.

4.8 Known Unknowns and Deep Uncertainties

While AMTAIR enhances reasoning under uncertainty, fundamental limitations remain regarding truly novel developments that might fall outside existing conceptual frameworks.

4.8.1 Categories of Deep Uncertainty

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

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

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

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

4.8.2 Adaptation Strategies for Deep Uncertainty

Rather than pretending to model the unmodelable, AMTAIR incorporates several strategies:

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

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

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

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

4.8.3 Robust Decision-Making Principles

Given deep uncertainty, certain decision principles become paramount:

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

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

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

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

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

4.9 Summary of Implications

The discussion reveals both the promise and limitations of computational approaches to AI governance coordination:

Technical Feasibility: Despite imperfections, automated extraction and formal modeling prove practically viable for complex AI risk arguments.

Epistemic Value: Making implicit models explicit, enabling systematic comparison, and supporting evidence integration enhance collective reasoning.

Practical Limitations: Extraction boundaries, false precision risks, and implementation dependencies require careful management.

Integration Potential: The approach complements rather than replaces existing governance frameworks, adding rigor without sacrificing flexibility.

Future Development: Technical enhancements, methodological extensions, and ecosystem growth could amplify impact.

Deep Uncertainty: Fundamental limits on predicting novel developments require maintaining humility and adaptability.

These findings suggest AMTAIR represents a valuable addition to the AI governance toolkit—not a panacea but a meaningful enhancement to our collective capacity for navigating unprecedented challenges.

5. Conclusion: Toward Coordinated AI Governance

Chapter Overview
Grade Weight: 10% | **Target Length**: ~14% of text (~4,200 words)
Requirements: Summarizes thesis and argument, outlines implications, notes limitations, points to future research

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, allows for unprecedented transparency and explainability of the entire process, 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 successfully processes complex arguments like Carlsmith's power-seeking AI model, extracting hierarchical structures and probability information.

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: Theoretical analysis suggests that formal comparison would reveal structural agreements across different expert worldviews even when probability estimates diverge—providing foundations for coordination.

Intervention Impacts: Policy evaluation capabilities demonstrate how formal models enable rigorous assessment of governance options. The ability to trace intervention effects through complex causal networks 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: The system struggles with implicit assumptions, complex conditionals, and ambiguous quantifiers. 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 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 causal models struggle to capture fully.

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 or political rhetoric 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

Bibliography

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 26, 2025

Valentin Meyer