Creating an Annotated Thesis Outline Based on Comprehensive Analysis

I've carefully analyzed all provided materials, including:

- Project documentation and descriptions
- Technical implementation files and code
- The detailed Colab notebook applying the AMTAIR approach to Carlsmith's model
- The University of Bayreuth's grading rubric

Based on this analysis, I'll now create a comprehensive thesis outline optimized to achieve the highest possible grade while effectively communicating the research contribution.

Automating Transformative AI Risk Modeling: A Computational Approach to Policy Impact Evaluation

Abstract [~300 words]

- Concise introduction to the coordination crisis in AI governance
- Brief explanation of the AMTAIR approach as a solution
- Summary of key innovations: automated extraction, BayesDown representation, interactive visualization
- Preview of application to Carlsmith's model and key findings
- Statement of research contribution to AI governance
- Note on implications for coordination across domains

1. Introduction [~2000 words, 10% of grade, ~ 14% of text]

1.1 The Coordination Crisis in AI Governance

- **Opening narrative:** Begin with concrete example of coordination failure in AI governance
- **Empirical paradox:** Juxtapose unprecedented investment with fundamental coordination gaps
- **Consequences:** Document systematic risk increases through safety gaps, resource misallocation, and negative-sum dynamics
- **Stakeholder mapping:** Analyze how technical researchers, policy specialists, and ethicists operate with different priorities and assumptions
- **Historical parallels:** Draw connections to nuclear governance, climate change, and biosecurity

- **Urgency factors:** Explain how accelerating capabilities compress available response time

1.2 Research Question and Scope

- **Primary question:** “How can frontier AI technologies be utilized to automate the extraction of probabilistic world models from AI safety literature, enabling robust prediction of policy impacts?”
- **Component definitions:** Define each element with precision: ‘frontier AI’, ‘automation’, ‘probabilistic world models’, ‘policy impacts’
- **Study boundaries:** Explicitly state scope limitations (focus on misaligned AI, not comprehensive governance)
- **Disciplinary positioning:** Situate the work at the intersection of AI safety, knowledge representation, and policy analysis
- **Approach justification:** Explain why computational approaches are needed for this particular challenge

1.3 The Multiplicative Benefits Framework

- **Core thesis:** Present the synergistic combination of (1) automated extraction, (2) prediction market integration, and (3) formal policy evaluation
- **Theoretical justification:** Explain how each component addresses specific epistemic challenges
- **Causal diagram:** Include visual representation of how components interact
- **Benefits explanation:** Provide concrete examples of multiplicative effects across domains

1.4 Thesis Structure and Roadmap

- **Overview of structure:** Preview the logical progression of the thesis
- **Linkage statements:** Explain how each section builds on previous ones
- **Signposting:** Create clear navigation guides for readers
- **Reading guidance:** Suggest different pathways for readers with different backgrounds

2. Background and Context [~4000 words, 20% of grade]

2.1 AI Existential Risk: The Carlsmith Model

- **Introduction to Carlsmith’s work:** Explain his structured approach to assessing existential risk

- **Six key premises:** Detail each premise with its original probability estimate
- **Composite risk calculation:** Show how Carlsmith derives ~5% probability
- **Significance:** Explain why this model represents an important contribution to AI risk assessment
- **Formalization potential:** Explain why this model is ideal for formal representation
- **CODE EXAMPLE:** Simple diagram showing Carlsmith's original probability calculation

2.2 Bayesian Networks as Knowledge Representation

- **Mathematical foundations:** Present formal definition and properties
- **DAG properties:** Explain directed acyclic graphs, nodes, edges, and conditional probability tables
- **RAIN-SPRINKLER-LAWN EXAMPLE:** Introduce this canonical example to illustrate key concepts
  - Include diagram showing the network structure
  - Present probability tables for each node
  - Walk through inference calculation examples
- **Cognitive advantages:** Explain why this formalism helps human reasoning about uncertainty
- **Application to AI risk:** Justify why Bayesian networks are particularly suited to this domain
- **CODE EXAMPLE:** Simple Python implementation of the Rain-Sprinkler-Lawn network

2.3 The Epistemic Challenge of Policy Evaluation

- **Unique difficulties:** Analyze challenges specific to AI governance policy evaluation
- **Traditional methods assessment:** Evaluate why established approaches fall short
- **Explicit representation requirements:** Establish necessary features for effective evaluation
- **Historical analogs:** Analyze partial parallels from nuclear policy, pandemic response, and climate governance
- **Innovation necessity:** Argue for novel approaches given AI's unique characteristics

2.4 Argument Mapping and Formal Representations

- **Conceptual bridge:** Position argument mapping as connection between natural language and formal models
- **Structural elements:** Detail components of argument maps

- **ArgDown introduction:** Present the structured syntax for argument representation
  - **CODE EXAMPLE:** Show basic ArgDown syntax highlighting hierarchical structure
  - **RAIN-SPRINKLER-LAWN EXAMPLE:** Demonstrate the canonical example in ArgDown format
- **BayesDown extension:** Explain how probabilistic information is incorporated
  - **CODE EXAMPLE:** Present BayesDown syntax with instantiations, priors, and posteriors
- **Transformation workflow:** Illustrate progression from natural language to structured representation

2.5 The MTAIR Framework: Achievements and Limitations

- **Project overview:** Present the Modeling Transformative AI Risks project's origins and approach
- **Key innovations:** Highlight the framework's contributions
- **Practical impact:** Discuss how MTAIR has influenced AI safety research
- **Limitation analysis:** Systematically examine constraints in the original approach
- **Automation potential:** Explain how these limitations motivate the current research

3. Own Position and Argument [~4000 words, 20% of grade, ~ 29% of text]

3.1 The AMTAIR Solution: Automation and Integration

- **Conceptual innovation:** Present AMTAIR as a computational extension of the MTAIR framework
- **Core insights:** Explain how automation addresses the key limitations of manual approaches
- **System architecture:** Overview of the pipeline from text to interactive models
- **Primary contributions:** Highlight the key innovations in the AMTAIR approach
- **Integration potential:** Discuss how the system connects with existing governance frameworks

3.2 The Two-Stage Extraction Process

- **Process overview:** Explain the separation of structure and probability extraction
- **Stage 1: Structure extraction**
  - **Process details:** Outline the steps for extracting argument structure

  - **CODE EXAMPLE:** Show key function for ArgDown parsing
  - **Visualization:** Demonstrate structural extraction for Carlsmith model

- **Stage 2: Probability integration**
  - **Process details:** Explain how probability information is incorporated
  - **Question generation:** Show how appropriate questions are derived from structure
  - **CODE EXAMPLE:** Show key function for BayesDown enhancement
  - **Visualization:** Demonstrate probability extraction for Carlsmith model

3.3 BayesDown: Bridging Qualitative and Quantitative Representation

- **Intermediate representation:** Explain the value of a hybrid representation
- **Syntax design principles:** Discuss the design considerations for BayesDown
- **Human readability:** Emphasize the importance of maintaining narrative connection
- **Machine processability:** Explain how the format enables computational analysis
- **CODE EXAMPLE:** Complete BayesDown representation of a simple argument
- **Preservation of context:** Discuss how BayesDown maintains important qualitative elements

3.4 Interactive Visualization and Exploration

- **Visualization challenges:** Discuss the difficulties in representing complex probabilistic models
- **Visual encoding principles:** Explain the approach to color, size, and interaction
- **User interaction design:** Detail the progressive disclosure of information
- **CODE EXAMPLE:** Key visualization function with HTML generation
- **Carlsmith model visualization:** Present and analyze the interactive representation
- **Cognitive benefits:** Explain how visualization enhances understanding of complex models

3.5 Beyond Extraction: Toward Policy Evaluation

- **Counterfactual analysis:** Explain how the system enables "what if" scenario exploration
- **Intervention modeling:** Discuss the approach to representing policy interventions
- **Cross-worldview comparison:** Explain how different perspectives can be formally compared
- **CODE EXAMPLE:** Simple intervention evaluation on Carlsmith model
- **Decision support framework:** Present the approach to supporting governance decisions
- **Integration with forecasting:** Outline the potential for live data incorporation

4. Implementation: The AMTAIR Prototype [~3000 words, 15% of grade, ~ 20% of text]

4.1 System Architecture and Data Flow

- **Component overview:** Present the five main system components
  - Text ingestion and preprocessing
  - LLM-powered extraction pipeline
  - Bayesian network construction
  - Visualization and interaction interface
  - Analysis and inference engine
- **Data flow diagram:** Visualize the progression from text to interactive model
- **Implementation technologies:** Detail the technical stack
- **Design principles:** Explain architectural choices
- **CODE EXAMPLE:** Show high-level module organization

4.2 The Rain-Sprinkler-Lawn Implementation

- **Example introduction:** Explain the canonical Bayesian network example
- **Stage 1: ArgDown representation**
  - **CODE EXAMPLE:** Show the ArgDown representation
  - **Process explanation:** Walk through the structural extraction process
- **Stage 2: BayesDown enhancement**
  - **CODE EXAMPLE:** Show the BayesDown representation
  - **Process explanation:** Walk through the probability extraction process
- **Stage 3: Bayesian network construction**
  - **CODE EXAMPLE:** Show the network construction code
  - **Visual result:** Present the visualization of the network
- **Inference demonstration:** Show conditional probability queries and results
- **Validation:** Compare computational results to analytical solutions

4.3 Application to Carlsmith's Model

- **Model complexity:** Discuss the scale and complexity of this real-world example
- **Extraction process:** Detail the steps taken to formalize Carlsmith's argument
- **Key parameters:** Present the critical probabilities and their interpretation
- **CODE EXAMPLE:** Show key extraction and processing steps

- **Structural analysis:** Examine the causal structure revealed by formalization
- **Influence analysis:** Identify the most significant factors affecting existential risk
- **Visual exploration:** Present interactive visualization of the complete model

4.4 Performance and Validation

- **Extraction quality metrics:** Evaluate the system's extraction accuracy
- **Performance benchmarks:** Present computational efficiency measurements
- **Expert validation:** Summarize feedback from domain experts
- **Limitation analysis:** Discuss current constraints and challenges
- **CODE EXAMPLE:** Validation code for extraction quality assessment
- **Error analysis:** Examine common failure modes and their implications

5. Analysis and Results [~3000 words, 15% of grade, ~ 20% of text]

5.1 Structural Insights from Carlsmith's Model

- **Graph analysis:** Present network metrics and their interpretation
- **Centrality measures:** Identify the most connected and influential nodes
- **Path analysis:** Examine critical pathways to existential catastrophe
- **Markov blanket analysis:** Identify minimal contextual information for key variables
- **CODE EXAMPLE:** Centrality calculation and interpretation code
- **Visual representation:** Show critical paths and nodes in the formalized model

5.2 Probabilistic Assessment and Sensitivity

- **Aggregate risk calculation:** Recompute Carlsmith's ~5% probability through the model
- **Sensitivity analysis:** Identify which parameters most significantly affect the outcome
- **Uncertainty propagation:** Examine how uncertainty in different nodes affects conclusions
- **CODE EXAMPLE:** Sensitivity analysis implementation
- **Risk factor ranking:** Present ordered list of risk factors by impact on outcome
- **Intervention potential:** Identify high-leverage intervention points

5.3 Policy Impact Evaluation

- **Intervention modeling:** Demonstrate how policy changes are represented in the model
- **Counterfactual analysis:** Present results of "what if" scenario exploration
- **Case study - Safety standards:** Evaluate impact of mandatory safety standards

- **Case study - Compute governance:** Evaluate impact of compute access restrictions
- **CODE EXAMPLE:** Policy intervention implementation
- **Robustness analysis:** Assess intervention effectiveness across parameter variations

5.4 Cross-Domain Integration Potential

- **Technical-policy bridge:** Assess how the approach connects technical and governance domains
- **Research prioritization insights:** Identify critical research areas based on model structure
- **Communication enhancement:** Evaluate improvements in cross-stakeholder understanding
- **Implementation pathways:** Suggest integration with existing governance frameworks
- **Adoption considerations:** Discuss factors affecting practical implementation
- **Future directions:** Outline potential extensions and applications

6. Counterclaims and Rebuttals [~2000 words, 10% of grade, ~ 14% of text]

6.1 Formalization Limitations

- **COUNTERCLAIM:** Present the argument that formal models oversimplify complex governance challenges
- **Supporting evidence:** Discuss examples where formalization has had negative consequences
- **REBUTTAL:** Argue that appropriate formalization enhances rather than replaces qualitative understanding
- **Evidence:** Present case studies where formal models improved governance
- **Synthesis:** Suggest a balanced approach that preserves important qualitative elements

6.2 Epistemic Humility Considerations

- **COUNTERCLAIM:** Discuss the risk of false precision and overconfidence in quantitative models
- **Supporting evidence:** Examine historical cases of model-induced overconfidence
- **REBUTTAL:** Explain how explicit representation of uncertainty enhances epistemic humility
- **Evidence:** Present research on how formalization can increase awareness of limitations
- **Synthesis:** Propose approaches to maintaining appropriate epistemic humility while formalizing

6.3 Democratic Governance Concerns

- **COUNTERCLAIM:** Present the argument that technical formalization may exclude stakeholders
- **Supporting evidence:** Discuss accessibility barriers and expertise requirements
- **REBUTTAL:** Argue that visualization and interactive exploration enhance rather than reduce accessibility
- **Evidence:** Present research on how interactive visualization improves stakeholder engagement
- **Synthesis:** Suggest design principles for ensuring inclusive access to formal models

6.4 Implementation Feasibility

- **COUNTERCLAIM:** Discuss practical challenges in scaling the approach to real governance contexts
- **Supporting evidence:** Examine resource requirements and institutional barriers
- **REBUTTAL:** Present incremental implementation paths with progressive enhancement
- **Evidence:** Provide examples of successful incremental adoption of formal methods
- **Synthesis:** Outline a realistic roadmap for incorporating formal models into governance

7. Conclusion and Outlook [~2000 words, 10% of grade, ~ 14% of text]

7.1 Summary of Key Contributions

- **Methodological innovation:** Recap the automated extraction approach
- **Technical achievements:** Summarize the implementation and its performance
- **Analytical insights:** Review key findings from applying the approach to Carlsmith's model
- **Governance implications:** Highlight the relevance for AI governance coordination
- **Integration potential:** Summarize how the approach connects diverse stakeholders

7.2 Limitations of the Current Implementation

- **Technical limitations:** Discuss extraction quality, computational constraints, and scalability
- **Conceptual limitations:** Examine simplifications and assumptions in the approach
- **Practical limitations:** Assess barriers to real-world implementation
- **Validation limitations:** Acknowledge constraints in the evaluation methodology
- **Ethical considerations:** Discuss potential unintended consequences

7.3 Future Research Directions

- **Technical enhancements:** Outline promising extensions to the extraction pipeline
- **Integration pathways:** Suggest connections with prediction markets and forecasting platforms
- **Application domains:** Identify other areas where the approach could be valuable
- **Long-term vision:** Present a roadmap for comprehensive AI governance modeling
- **Research agenda:** Propose specific research questions for further investigation

7.4 Broader Implications for AI Governance

- **Epistemic infrastructure:** Discuss how formal modeling enhances community knowledge
- **Coordination mechanisms:** Examine how shared representations facilitate collaboration
- **Strategic planning:** Explore applications to long-term governance strategy
- **Institutional design:** Suggest governance structures that incorporate formal modeling
- **Normative reflections:** Consider the ethical dimensions of formalized risk assessment

8. References

- Full bibliography organized by topic area
- Primary sources for AI safety and governance literature
- Technical references for Bayesian networks and computational methods
- Sources for the Carlsmith model and other risk assessments
- Methodological references for formal modeling in governance contexts

Appendices

Appendix A: Technical Implementation Details

- **Environment setup:** Detailed software requirements and configuration
- **Full code listings:** Complete implementation of the extraction pipeline
- **API specifications:** Documentation of interfaces for each component
- **Data format specifications:** Detailed structure definitions
- **Development workflow:** Implementation process documentation

Appendix B: BayesDown Syntax Specification

- **Core syntax rules:** Comprehensive specification of the BayesDown syntax
- **Grammar definition:** Formal grammar in Extended Backus-Naur Form
- **Validation rules:** Specifications for checking well-formedness
- **Extension mechanisms:** Guidelines for syntax extensions
- **Migration guidelines:** Converting between different representation formats

Appendix C: Complete Carlsmith Model Analysis

- **Full model specification:** Complete BayesDown representation
- **Parameter derivation:** Explanation of how probabilities were determined
- **Comprehensive results:** Complete analysis outputs
- **Alternative interpretations:** Exploration of different model formulations
- **Expert feedback:** Documentation of validation with domain experts

Appendix D: Additional Case Studies

- **Alternative risk models:** Application to other AI risk frameworks
- **Real-world policy scenarios:** Evaluation of proposed governance mechanisms
- **Comparison with manual analysis:** Side-by-side comparison with traditional approaches
- **User study results:** Documentation of how stakeholders interact with the system
- **Extended validation:** Additional performance and accuracy assessments

Thesis Writing Journey Begins

Abstract

The coordination crisis in AI governance presents a paradoxical challenge: unprecedented investment in AI safety coexists alongside fundamental coordination failures across technical, policy, and ethical domains. These divisions systematically increase existential risk by creating safety gaps, misallocating resources, and fostering inconsistent approaches to interdependent problems. This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach that addresses this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models.

The AMTAIR system implements an end-to-end pipeline that transforms unstructured text into interactive Bayesian networks through a novel two-stage extraction process: first capturing argument structure in ArgDown format, then enhancing it with probability information in

BayesDown format. This approach bridges qualitative expert reasoning with quantitative analysis, making implicit models explicit and enabling rigorous evaluation of policy impacts. When applied to Joseph Carlsmith's model of existential risk from power-seeking AI, the system successfully formalizes complex causal relationships while preserving key narrative elements, revealing critical risk pathways and intervention opportunities.

By making implicit models explicit, enabling cross-worldview comparison, and supporting policy evaluation across diverse scenarios, the AMTAIR approach creates epistemic infrastructure that facilitates coordination between technical, governance, and ethical domains. This research offers both methodological innovations in automated knowledge extraction and practical tools for enhancing strategic coordination in AI governance—a critical contribution as capabilities continue to accelerate and the window for establishing effective governance narrows.

1. Introduction

1.1 The Coordination Crisis in AI Governance

On March 22, 2023, over 1,000 AI researchers and technology leaders signed an open letter calling for a pause in advanced AI development, citing "profound risks to society and humanity." Within days, multiple counterstatements emerged—some arguing the risks were overstated, others that the proposed pause was insufficient, and still others that the entire framing misunderstood the problem. This fragmentation of response typifies what I call the coordination crisis in AI governance: despite unprecedented investment and growing awareness, we lack the strategic "operating system" needed to align disparate efforts as AI capabilities advance at an accelerating pace.

This coordination gap isn't merely inefficient—it systematically increases existential risk. When organizations function as independent processors without shared protocols, we generate duplicative work, leave critical gaps unaddressed, and create inconsistent approaches to interdependent problems. Technical alignment researchers develop solutions without implementation pathways; policy specialists craft frameworks without technical grounding; ethicists articulate principles without operational specificity. As capabilities approach human-level intelligence, this fragmentation becomes increasingly dangerous.

The empirical patterns defining this landscape reveal troubling trends. First, AI capabilities are advancing at an accelerating pace, with compression from decades to months between significant milestones and emergent capabilities appearing at scale thresholds. Second, technical alignment efforts face substantial challenges, including specification problems, robustness limitations, and interpretability bottlenecks. Third, AI governance efforts remain fragmented with proliferation without convergence, institutional silos, and competing jurisdictional claims. Finally, global coordination mechanisms have consistently struggled with analogous challenges from climate change to nuclear security to pandemic response, suggesting existing institutions are poorly suited to rapid technological development with distributed creation capability.

Historical parallels highlight the unique difficulties in the AI domain. Early nuclear governance relied on implicit coordination with devastating consequences; only after explicit mechanisms emerged—test ban treaties, verification protocols—did risks stabilize. Similarly, climate change coordination suffered decades of delay when lacking shared models and verification mechanisms. What distinguishes AI governance, however, is the compressed timeframe for action, the technical complexity requiring integration across disciplines, and the mixed competitive-cooperative incentives that create classic stag hunt dynamics with tragedy-of-the-commons characteristics.

This coordination crisis demands novel approaches to knowledge sharing and integration across domains. As capabilities accelerate and the window for establishing effective governance narrows, better tools for facilitating coordination become not merely beneficial but essential for managing what may be humanity's most consequential technological challenge.

1.2 Research Question and Scope

This thesis addresses a specific aspect of the coordination crisis in AI governance through the central research question: **How can frontier AI technologies be utilized to automate the extraction of probabilistic world models from AI safety literature, enabling robust prediction of policy impacts?**

To properly frame this investigation, I must clearly define the key components of this question:

Frontier AI technologies refers to the most capable large language models (LLMs) and related systems that demonstrate advanced capabilities in understanding and generating text, analyzing complex patterns, and performing structured transformations of information. These technologies serve both as the subject of governance concern and, in this research, as tools for addressing governance challenges.

Automation involves creating computational systems that can perform tasks previously requiring human expertise with minimal supervision, particularly the extraction of structured representations from unstructured text and the transformation of these representations into formal models.

Probabilistic world models are formalized representations of causal relationships and uncertainties that capture both the structure of arguments (which factors influence which outcomes) and quantitative judgments about likelihoods (how probable different scenarios are based on various conditions). These models make implicit reasoning explicit and enable rigorous analysis.

Policy impacts refers to the counterfactual effects of governance interventions on outcomes of interest, particularly the reduction of existential risk from advanced AI systems. Predicting these impacts involves modeling how changes in relevant factors (such as safety standards, development practices, or coordination mechanisms) affect the probabilities of different scenarios.

The scope of this research is carefully bounded in several important ways. First, it focuses specifically on existential risk from misaligned AI rather than attempting to address all AI governance challenges. Second, it examines automation of existing expert knowledge rather than generating novel risk assessments. Third, it prioritizes making implicit models explicit rather than advocating for particular governance positions. Finally, it emphasizes the extraction and representation of arguments rather than developing novel infrastructure for forecasting and prediction markets, though it enables integration with such systems.

This research sits at the intersection of several disciplines, drawing on technical AI alignment (for understanding risk factors), knowledge representation (for formal modeling approaches), and AI governance (for policy context and intervention options). It employs computational methods not as a replacement for human judgment but as tools to enhance the accessibility, precision, and integration of expert reasoning across domains. This hybrid approach acknowledges both the technical complexity of AI risk assessment and the inherently value-laden nature of governance decisions.

1.3 The Multiplicative Benefits Framework

The approach developed in this thesis combines three complementary elements that, when integrated, create value that exceeds their individual contributions. This multiplicative benefits framework explains why the components must be developed together rather than separately:

First, **automated extraction** transforms unstructured expert knowledge into structured representations, making implicit models explicit and enabling rigorous analysis. While valuable on its own, extraction reaches its full potential when the resulting models are enhanced with probabilistic information and connected to live data sources. The extraction component addresses the key challenge of scaling up formalization beyond what manual approaches can achieve, using frontier LLMs to process the growing volume of AI safety literature.

Second, **prediction market integration** connects static models to dynamic data streams, ensuring that risk assessments remain current as new information emerges. This component bridges the gap between theoretical frameworks and empirical evidence, creating living models that evolve with the rapidly changing AI landscape. While prediction markets provide valuable information independently, their integration with formal causal models dramatically enhances their utility for understanding complex risk scenarios.

Third, **formal policy evaluation** enables rigorous assessment of governance interventions, testing how specific proposals might perform across different possible futures. This component transforms abstract policy discussions into concrete, quantifiable assessments of expected impact, helping governance stakeholders allocate resources to the most effective interventions.

While policy analysis can be conducted without formal models, the ability to systematically evaluate interventions across diverse worldviews substantially improves analysis quality.

These components interact in synergistic ways illustrated by the causal diagram in Figure 1. Automated extraction provides the foundation by transforming unstructured knowledge into formal models. These models then serve as the structure for integrating prediction market data, which updates the probability estimates. The enhanced models enable formal policy evaluation, which generates insights that inform both the models themselves and real-world governance decisions.

[FIGURE 1: Causal diagram showing interactions between automated extraction, prediction market integration, and policy evaluation components]

Consider a concrete example of these multiplicative benefits: when analyzing proposals for governance of compute resources, automated extraction might formalize expert perspectives on how compute access affects capability development and risk. Prediction market integration could then provide current estimates of key uncertainties like technological development timelines. Policy evaluation would use this enhanced model to compare different compute governance approaches across various scenarios, revealing which approaches remain robust despite uncertainty about future developments.

Without any one component, the system's value would be substantially diminished. Extraction without prediction markets would create static models that quickly become outdated. Prediction markets without formal causal models would provide isolated data points without coherent integration. Policy evaluation without automated extraction would be limited to a small set of manually created models, missing the diversity of expert perspectives. The full value emerges only when all components work together to create a comprehensive system for understanding and managing AI risk.

1.4 Thesis Structure and Roadmap

This thesis proceeds through a structured progression designed to build a comprehensive understanding of both the coordination challenge and the proposed solution. Each section builds upon previous ones while addressing specific aspects of the research question.

In **Section 2: Background and Context**, I establish the theoretical and practical foundations for the research. First, I introduce Carlsmith's model of existential risk from power-seeking AI, explaining its structured approach to quantifying risk through six key premises. Then I examine Bayesian networks as a knowledge representation framework, using the canonical rain-sprinkler-lawn example to illustrate fundamental concepts. Next, I analyze the unique epistemic challenges in policy evaluation for AI governance, explaining why traditional approaches fall short. Finally, I explore argument mapping and formal representations as bridges between qualitative reasoning and quantitative models, introducing the ArgDown and BayesDown formats.

Section 3: Own Position and Argument presents the AMTAIR approach as a solution to the coordination crisis. I explain the system architecture, with particular focus on the two-stage extraction process that separates structure from probability. I then explore BayesDown as a hybrid representation bridging qualitative and quantitative aspects. Next, I discuss the interactive visualization approach that makes complex models accessible to diverse stakeholders. Finally, I outline how the system enables policy evaluation through counterfactual analysis and intervention modeling.

In **Section 4: Implementation**, I detail the technical realization of the AMTAIR approach. Beginning with the system architecture and data flow, I explain how components interact to transform text into interactive models. I then demonstrate the complete pipeline using the canonical rain-sprinkler-lawn example, walking through each stage of the process with code examples and visualizations. Next, I apply the system to Carlsmith's model, showing how a complex real-world risk assessment can be formalized and analyzed. Finally, I present performance metrics and validation results demonstrating the system's capabilities and limitations.

Section 5: Analysis and Results examines insights gained from applying the AMTAIR approach to Carlsmith's model. I analyze structural properties of the formalized model, including centrality measures and critical pathways. I then perform sensitivity analysis to identify the most influential parameters affecting risk estimates. Next, I demonstrate policy impact evaluation by modeling specific interventions and assessing their effects across scenarios. Finally, I discuss cross-domain integration potential, examining how the approach can connect technical, governance, and ethical domains.

In **Section 6: Counterclaims and Rebuttals**, I address potential objections to the AMTAIR approach. I examine limitations of formalization, concerns about epistemic humility, democratic governance considerations, and implementation feasibility challenges. For each objection, I present supporting evidence, offer a rebuttal, and suggest a synthesis that acknowledges the valid concerns while demonstrating how the approach addresses them.

Section 7: Conclusion and Outlook summarizes key contributions, acknowledges limitations, and explores future research directions. I recap methodological innovations, technical achievements, and analytical insights before discussing remaining challenges. I then outline promising extensions to the system and suggest broader applications. Finally, I reflect on implications for AI governance, discussing how formal modeling can enhance epistemics, facilitate coordination, and inform strategic planning.

The thesis includes comprehensive **References** and **Appendices** with technical details, syntax specifications, complete analysis results, and additional case studies.

Readers with technical backgrounds may wish to focus initially on Sections 4 and 5, which provide detailed implementation information and results. Those primarily interested in AI governance may find Sections 3 and 6 most relevant to policy considerations. For readers new to the topic, following the sections in sequence will build a progressive understanding from foundational concepts to specific applications and implications.

2. Background and Context

2.1 AI Existential Risk: The Carlsmith Model

Joseph Carlsmith's "Is Power-Seeking AI an Existential Risk?" represents one of the most structured attempts to assess the probability of existential catastrophe from advanced AI systems. Rather than relying on intuition or general concerns, Carlsmith approaches the question by breaking it down into six key premises with explicitly estimated probabilities. This decomposition makes his model an ideal candidate for formalization, as it already exhibits a structure amenable to Bayesian network representation.

Carlsmith's six key premises, each with his probability estimates, are:

1. **Transformative AI this century (80%)**: "By 2100, humans will develop AI systems that can perform almost all economically relevant human cognitive labor much more cheaply than humans."
2. **AI systems pursuing objectives (95%)**: "If we develop TAI systems, we will build and deploy systems that pursue objectives in the world."
3. **Systems with power-seeking incentives (40%)**: "Some of these systems will have objectives and capabilities that create strong incentives for power-seeking behavior."
4. **Systems with sufficient capability for existential threat (65%)**: "Power-seeking systems of this kind will have strong capability advantages over humans."
5. **Misaligned systems (50%)**: "Some of these systems will be goal-misaligned with the continued existence of humans."
6. **Misaligned power-seeking systems causing catastrophe (65%)**: "Efforts to create aligned and safe systems will fall short in critical cases."

By multiplying these probabilities (with some adjustments for dependencies), Carlsmith arrives at an approximately 5% probability of existential catastrophe from power-seeking AI. This estimate represents his considered judgment after extensive research and consultation with domain experts.

What makes Carlsmith's model particularly valuable for formal representation is not just its explicit probabilities, but its clearly articulated causal structure. He describes how these premises connect and influence one another, creating a framework that naturally translates into a Bayesian network. For example, he explains how the difficulty of alignment influences the likelihood of misaligned systems, and how various factors might enable or prevent catastrophic outcomes from misaligned systems.

The model goes beyond these six premises to explore additional factors. Carlsmith discusses how instrumental convergence, problems with proxies, and problems with search processes contribute to the difficulty of alignment. He examines how warning shots, rapid capability escalation, and corrective feedback affect the likelihood of societal responses. He considers incentives to deploy potentially dangerous systems and deception by AI systems as important factors in deployment decisions.

Figure 2 shows a simplified diagram of Carlsmith's model, highlighting the key causal relationships between factors.

[FIGURE 2: Simplified diagram of Carlsmith's model showing causal relationships between key factors]

```
# Simple code to calculate Carlsmith's bottom-line probability
p_transformative_ai = 0.8
p_objective_pursuit = 0.95
p_power_seeking = 0.4
p_capability_advantage = 0.65
p_misalignment = 0.5
p_catastrophe_given_all_above = 0.65

# Simplified calculation (ignoring some dependencies)
p_doom = (p_transformative_ai * p_objective_pursuit * p_power_seeking *
          p_capability_advantage * p_misalignment * p_catastrophe_given_all_above)

print(f"Estimated probability of existential catastrophe: {p_doom:.3f} or about {p_doom*100:
```

While this calculation provides a useful starting point, it simplifies important dependencies between the factors. For example, the likelihood of catastrophe given misaligned power-seeking systems is not independent of the capability advantage those systems have. A more sophisticated model needs to represent these conditional dependencies explicitly—precisely what a Bayesian network approach enables.

Carlsmith's model provides an ideal case study for the AMTAIR approach for several reasons. First, it contains explicit probability estimates that can be captured in a formal representation. Second, it has a clear causal structure linking various factors that contribute to risk. Third, it encompasses a wide range of considerations from technical alignment to governance factors, making it relevant across domains. Finally, its structure is complex enough to demonstrate the value of formalization while remaining tractable for analysis.

By formalizing Carlsmith's model, we can not only preserve his original analysis but enhance it through structural examination, sensitivity analysis, and policy evaluation—tasks that become possible once the implicit model is made explicit through computational representation.

2.2 Bayesian Networks as Knowledge Representation

Bayesian networks provide a powerful framework for representing and reasoning about uncertain knowledge, making them particularly suitable for modeling complex domains like AI risk. A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).

Formally, a Bayesian network consists of:

1. A set of variables {X , X , …, X } representing different aspects of the domain
2. A directed acyclic graph where nodes represent variables and edges represent direct dependencies
3. A conditional probability distribution P(X |Parents(X)) for each variable X

The network structure encodes conditional independence assumptions: each variable X is conditionally independent of its non-descendants given its parents in the graph. This property enables compact representation of joint probability distributions, which would otherwise require exponentially many parameters.

The canonical "Rain-Sprinkler-Lawn" example illustrates these concepts simply but effectively. Consider a scenario with three binary variables:

- Rain (R): Whether it is raining (TRUE/FALSE)
- Sprinkler (S): Whether the sprinkler is on (TRUE/FALSE)
- Grass_Wet (W): Whether the grass is wet (TRUE/FALSE)

Both rain and the sprinkler can cause the grass to be wet, and rain also influences whether the sprinkler is on (people typically don't use sprinklers when it's raining). Figure 3 shows this network structure.

[FIGURE 3: Diagram of the Rain-Sprinkler-Lawn Bayesian network showing Rain influencing both Sprinkler and Grass_Wet, and Sprinkler influencing Grass_Wet]

For each node, we specify a conditional probability table (CPT) defining the probability distribution over its possible values, conditioned on all possible combinations of its parent values. For example:

- P(R=TRUE) = 0.2, P(R=FALSE) = 0.8 (prior probability of rain)
- P(S=TRUE|R=TRUE) = 0.01, P(S=TRUE|R=FALSE) = 0.4 (conditional probability of sprinkler given rain)
- P(W=TRUE|R=TRUE,S=TRUE) = 0.99, P(W=TRUE|R=TRUE,S=FALSE) = 0.8, etc. (conditional probability of wet grass given rain and sprinkler)

With this representation, we can compute the probability of any combination of variable values or answer queries about conditional probabilities. For example, we can calculate the probability that it was raining given that the grass is wet, P(R=TRUE|W=TRUE), using Bayes' rule and the conditional probabilities in the network.

```
import numpy as np
import networkx as nx
import matplotlib.pyplot as plt
from pgmpy.models import BayesianNetwork
from pgmpy.factors.discrete import TabularCPD

# Define the network structure
model = BayesianNetwork([('R', 'S'), ('R', 'W'), ('S', 'W')])

# Define conditional probability distributions
cpd_r = TabularCPD(variable='R', variable_card=2, values=[[0.2], [0.8]],
                  state_names={'R': ['TRUE', 'FALSE']})

cpd_s = TabularCPD(variable='S', variable_card=2,
                  values=[[0.01, 0.4], [0.99, 0.6]],
                  evidence=['R'], evidence_card=[2],
                  state_names={'S': ['TRUE', 'FALSE'], 'R': ['TRUE', 'FALSE']})

cpd_w = TabularCPD(variable='W', variable_card=2,
                  values=[[0.99, 0.8, 0.9, 0.0], [0.01, 0.2, 0.1, 1.0]],
                  evidence=['R', 'S'], evidence_card=[2, 2],
                  state_names={'W': ['TRUE', 'FALSE'], 'R': ['TRUE', 'FALSE'], 'S': ['TRUE',

# Add CPDs to the model
model.add_cpds(cpd_r, cpd_s, cpd_w)

# Check model validity
model.check_model()

# Create visual representation
G = nx.DiGraph()
G.add_edges_from([('R', 'S'), ('R', 'W'), ('S', 'W')])
pos = {'R': (0, 1), 'S': (1, 1), 'W': (0.5, 0)}

plt.figure(figsize=(8, 6))
nx.draw(G, pos, with_labels=True, node_size=3000, node_color='lightblue',
       font_size=12, font_weight='bold', arrowsize=20)
plt.title('Rain-Sprinkler-Lawn Bayesian Network')
plt.show()

# Example inference: P(Rain=TRUE | Grass_Wet=TRUE)
from pgmpy.inference import VariableElimination
inference = VariableElimination(model)
result = inference.query(variables=['R'], evidence={'W': 'TRUE'})
print("P(Rain=TRUE | Grass_Wet=TRUE) =", result.values[0])
```

Bayesian networks offer several advantages for modeling AI risk:

1. **Causal interpretation:** The directed edges represent causal influences, aligning with our natural understanding of how factors affect outcomes.

2. **Uncertainty representation:** They explicitly represent probability distributions, capturing the inherent uncertainty in complex domains.

3. **Modular structure:** New variables and relationships can be added without rebuilding the entire model, enabling incremental refinement.

4. **Inference capability:** They support various types of queries, including prediction (what will happen given current conditions?), diagnosis (what might have caused observed outcomes?), and intervention (what if we change something?).

5. **Transparency:** The structure and parameters are explicitly defined, making assumptions and judgments transparent for critique and refinement.

Perhaps most importantly, Bayesian networks align with how human experts often think about complex problems: identifying key factors, understanding how they influence each other, and making judgments about likelihoods under different conditions. This makes them well-suited for representing expert knowledge in a format that supports both human understanding and computational analysis.

The Rain-Sprinkler-Lawn example, while simple, illustrates the core concepts we'll apply to much more complex domains like AI risk. The same principles of identifying variables, specifying their relationships, and quantifying conditional probabilities extend naturally to models with dozens or hundreds of variables representing the many factors that influence existential risk from advanced AI systems.

2.3 The Epistemic Challenge of Policy Evaluation

Evaluating policy interventions for AI governance presents unique epistemic challenges that traditional policy analysis methods struggle to address. These challenges arise from the complex causal chains, deep uncertainty, divergent worldviews, and limited empirical grounding that characterize the domain.

Traditional policy analysis relies heavily on historical precedent, empirical data, and established causal models. Cost-benefit analysis quantifies the predicted impacts of interventions based on observed relationships between variables. Scenario planning explores different futures but typically lacks probability estimates. Expert elicitation captures specialist knowledge but often fails to systematically represent interdependencies between factors. None of these approaches fully addresses the specific challenges of AI governance policy evaluation.

Four unique difficulties define the epistemic landscape of AI governance:

First, **complex causal chains with limited empirical grounding** characterize the relationship between governance interventions and risk outcomes. Unlike domains like public health, where interventions have measurable effects on well-defined outcomes, AI governance involves extended causal chains where actions today might influence technological development paths, institutional behaviors, and ultimately risk profiles decades in the future. These chains cannot be empirically tested through traditional methods, yet understanding them is essential for effective governance.

Second, **deep uncertainty about future capability development** creates a challenging environment for prediction. While some aspects of technology evolution follow discernible patterns, transformative capabilities often emerge unexpectedly through conceptual breakthroughs. This uncertainty isn't merely quantitative (what are the error bars on our predictions?) but qualitative (what kinds of capabilities might emerge?), creating fundamental challenges for traditional forecasting methods that rely on extrapolation from past trends.

Third, **divergent worldviews about fundamental risk factors** complicate consensus-building around governance approaches. Experts disagree not just about probability estimates but about which factors matter most and how they relate causally. Some emphasize technical alignment challenges, others focus on competitive dynamics between developers, and still others prioritize institutional oversight mechanisms. Each worldview implies different intervention priorities, yet traditional policy analysis lacks tools for systematically comparing perspectives.

Fourth, **limited opportunities for experimental testing** prevent iterative refinement of governance approaches. Unlike domains where small-scale pilots can test intervention efficacy before wider implementation, many AI governance interventions must be designed without the benefit of experimental evidence. If certain risks materialize only once systems reach advanced capabilities, learning from experience comes too late.

Addressing these challenges requires explicit representation across multiple dimensions:

- **Uncertainty across multiple parameters:** The approach must represent not just uncertainty about outcomes but uncertainty about the relationships between variables and the structure of the causal model itself.

- **Conditional dependencies between variables:** The system needs to capture how different factors influence each other, enabling understanding of complex chains of causation from interventions to outcomes.
- **Comparable representation of different worldviews:** To facilitate productive discourse across perspectives, the approach must represent diverse causal models in a common framework that highlights both agreements and disagreements.
- **Continuous evidence integration mechanisms:** As new information emerges—from theoretical insights, empirical observations, or expert judgments—the system should update its representations to reflect current knowledge.

Historical analogues provide partial insights but no complete template. Nuclear governance established verification protocols and international monitoring, but over a longer timeframe than likely available for AI. Pandemic response developed early warning systems and response protocols, but struggles with similar challenges in predicting novel pathogen emergence. Climate governance demonstrates the difficulty of establishing effective international coordination mechanisms for slow-moving, high-impact risks.

What distinguishes AI governance is the combination of accelerating technological development, distributed creation capability, and potentially irreversible consequences once certain thresholds are crossed. This unique profile necessitates novel approaches to policy evaluation that can handle the epistemic challenges described above while providing actionable insights for governance.

The formal modeling approach developed in this thesis addresses these challenges by making assumptions explicit, facilitating structured comparison of worldviews, and enabling rigorous exploration of intervention impacts across scenarios. By transforming implicit models into explicit representations, it creates a foundation for more productive discourse about governance priorities and approaches, even amid deep uncertainty about future developments.

2.4 Argument Mapping and Formal Representations

Argument mapping provides a bridge between natural language reasoning and formal probabilistic models, enabling the transformation of complex qualitative arguments into structured representations suitable for computational analysis. This section explores two key intermediate representations—ArgDown and BayesDown—that facilitate this transformation process.

Argument maps are structured visualizations that represent the logical relationships between claims, evidence, and objections. Unlike free-form text, they make explicit how different statements support or challenge one another, forcing clarity about the logical structure of arguments. Traditional argument maps typically include:

- Statements (claims, premises, conclusions) presented as nodes
- Support and attack relationships shown as arrows between nodes

- Hierarchical organization reflecting logical dependencies

These visualizations help identify unstated assumptions, circular reasoning, and gaps in argumentation. However, traditional argument mapping has limited expressivity for representing uncertainty—a crucial element in complex domains like AI risk assessment.

ArgDown extends the concept of argument mapping into a structured text format with a consistent syntax. Developed by Christian Voigt at Karlsruhe Institute of Technology, ArgDown provides a markdown-like notation for representing arguments in a hierarchical structure that can be automatically visualized and analyzed. The basic syntax is:

```
[Statement]: Description of the statement.
 + [Supporting_Statement]: Description of supporting statement.
   + [Further_Support]: Description of additional support.
 - [Opposing_Statement]: Description of opposing statement.
```

For the AMTAIR project, we adapt ArgDown to focus on causal relationships rather than general argumentation, using a modified syntax where the hierarchical structure represents causal influence:

```
[Effect]: Description of effect. {"instantiations": ["effect_TRUE", "effect_FALSE"]}
 + [Cause1]: Description of first cause. {"instantiations": ["cause1_TRUE", "cause1_FALSE"]}
 + [Cause2]: Description of second cause. {"instantiations": ["cause2_TRUE", "cause2_FALSE"]
   + [Root_Cause]: A cause that influences Cause2. {"instantiations": ["root_TRUE", "root_FA
```

This adaptation adds metadata in JSON format to specify possible states (instantiations) of each variable, preparing the structure for probabilistic enhancement. The hierarchical relationships (indented with plus signs) represent causal influence, creating a directed graph structure.

The Rain-Sprinkler-Lawn example in ArgDown format illustrates this structure:

```
[Grass_Wet]: Concentrated moisture on, between and around the blades of grass. {"instantiatio
 + [Rain]: Tears of angles crying high up in the skies hitting the ground. {"instantiations"
 + [Sprinkler]: Activation of a centrifugal force based CO2 droplet distribution system. {"in
   + [Rain]
```

This representation captures the causal structure (both Rain and Sprinkler influence Grass_Wet, and Rain also influences Sprinkler) and specifies the possible states of each variable. However, it lacks probability information, which is where BayesDown extends the representation.

BayesDown builds on ArgDown by adding probability metadata, transforming a purely structural representation into a complete Bayesian network specification. The enhanced format includes:

```
[Node]: Description. {
  "instantiations": ["node_TRUE", "node_FALSE"],
  "priors": {
    "p(node_TRUE)": "0.7",
    "p(node_FALSE)": "0.3"
  },
  "posteriors": {
    "p(node_TRUE|parent_TRUE)": "0.9",
    "p(node_TRUE|parent_FALSE)": "0.4",
    "p(node_FALSE|parent_TRUE)": "0.1",
    "p(node_FALSE|parent_FALSE)": "0.6"
  }
}
```

The Rain-Sprinkler-Lawn example in BayesDown format illustrates this enhancement:

```
[Grass_Wet]: Concentrated moisture on grass. {
  "instantiations": ["grass_wet_TRUE", "grass_wet_FALSE"],
  "priors": {"p(grass_wet_TRUE)": "0.322", "p(grass_wet_FALSE)": "0.678"},
  "posteriors": {
    "p(grass_wet_TRUE|sprinkler_TRUE,rain_TRUE)": "0.99",
    "p(grass_wet_TRUE|sprinkler_TRUE,rain_FALSE)": "0.9",
    "p(grass_wet_TRUE|sprinkler_FALSE,rain_TRUE)": "0.8",
    "p(grass_wet_TRUE|sprinkler_FALSE,rain_FALSE)": "0.0"
  }
}
 + [Rain]: Water falling from the sky. {
   "instantiations": ["rain_TRUE", "rain_FALSE"],
   "priors": {"p(rain_TRUE)": "0.2", "p(rain_FALSE)": "0.8"}
 }
 + [Sprinkler]: Artificial watering system. {
   "instantiations": ["sprinkler_TRUE", "sprinkler_FALSE"],
   "priors": {"p(sprinkler_TRUE)": "0.44838", "p(sprinkler_FALSE)": "0.55162"},
   "posteriors": {
     "p(sprinkler_TRUE|rain_TRUE)": "0.01",
     "p(sprinkler_TRUE|rain_FALSE)": "0.4"
   }
 }
```

```
  + [Rain]
```

This representation now contains all the information needed to construct a complete Bayesian network: variables with their possible states, causal relationships between variables, prior probabilities for root nodes, and conditional probability tables for nodes with parents.

The transformation workflow from natural language to BayesDown involves several steps:

1. Identify key variables and their possible states from the text
2. Determine causal relationships between variables
3. Represent the structure in ArgDown format
4. Generate probability questions based on the structure
5. Answer these questions (manually or via LLM)
6. Incorporate probability answers into BayesDown format

This progressive transformation preserves the narrative richness of the original text while adding formal structure. The intermediate representations (ArgDown and BayesDown) remain human-readable, maintaining the connection to the original arguments while enabling computational analysis.

The key innovation in this approach is the separation of structure extraction from probability quantification, which aligns with how experts typically approach complex arguments. First, they identify what factors matter and how they relate causally, then they consider how probable different scenarios are based on those relationships. This two-stage process makes the extraction more robust and the resulting representations more interpretable.

2.5 The MTAIR Framework: Achievements and Limitations

The Modeling Transformative AI Risks (MTAIR) project, led by David Manheim and colleagues, represents a significant precursor to the current research. Launched in 2021, MTAIR aimed to create structured representations of existential risks from advanced AI using Bayesian networks, directed acyclic graphs, and probabilistic modeling. Understanding its achievements and limitations provides important context for the current AMTAIR approach.

MTAIR emerged from the recognition that AI risk discussions often involved complex causal arguments with implicit probability judgments that were difficult to compare or integrate. By formalizing these arguments in structured models, the project sought to make assumptions explicit, enable quantitative analysis, and facilitate more productive discourse across different perspectives on AI risk.

The framework's key innovations included:

1. **Explicit representation of uncertainty through probability distributions:** Rather than presenting point estimates, MTAIR captured uncertainty about parameters using distributions, acknowledging the significant uncertainty in AI risk assessment.

2. **Hierarchical structure for complex scenarios:** The approach used nested models that allowed exploration of different levels of detail, from high-level risk factors to specific technical mechanisms.
3. **Integration of diverse expert judgments:** The framework incorporated perspectives from various specialists, creating a more comprehensive view than any single expert could provide.
4. **Sensitivity analysis methodology:** MTAIR developed techniques for identifying which parameters most significantly affected risk estimates, helping prioritize research efforts.

The project's practical impact extended beyond its technical achievements. It influenced research prioritization by identifying critical uncertainties that warranted further investigation. It enhanced discourse quality by providing a shared vocabulary and structure for discussing causal pathways to risk. It also created visual representations that made complex arguments more accessible to stakeholders without technical backgrounds.

Despite these achievements, MTAIR faced several important limitations:

1. **Manual labor intensity limiting scalability:** Creating and updating models required substantial expert time, limiting the number and complexity of models that could be developed and maintained. As one team member noted, "It often took several days of work to formalize even relatively straightforward arguments."
2. **Static nature of models once constructed:** The models were essentially snapshots that did not automatically update as new information emerged, requiring manual revision to remain current.
3. **Limited accessibility for non-technical stakeholders:** While visual representations improved accessibility, understanding and interacting with the models still required specialized knowledge.
4. **Challenges in representing multiple worldviews simultaneously:** Comparing different perspectives required creating separate models, making it difficult to identify specific points of agreement and disagreement.

These limitations motivate the current research in automating the extraction and transformation process. As AI capabilities advance and the volume of relevant research grows, manual approaches cannot keep pace with the need for comprehensive, up-to-date models. Automation addresses the scalability limitation by dramatically reducing the time required to create formal representations of expert arguments.

Moreover, incorporating frontier LLMs into the pipeline enables new capabilities that were not feasible in the original MTAIR framework. These include:

1. Processing larger volumes of literature to capture more diverse perspectives

2. Generating intermediate representations that preserve narrative structure
3. Automating the creation of probability questions based on model structure
4. Facilitating integration with live data sources for continuous updates

By building on MTAIR's foundation while addressing its key limitations, the current research maintains continuity with established approaches to AI risk modeling while pushing the boundaries of what's possible through automation and enhanced representation formats.

The evolution from MTAIR to AMTAIR represents a natural progression: as the field matures and the challenges become more pressing, more sophisticated tools are needed to facilitate coordination and decision-making. Automation doesn't replace expert judgment but amplifies it, allowing insights to be captured, formalized, and shared more efficiently across the AI governance community.

3. Own Position and Argument

3.1 The AMTAIR Solution: Automation and Integration

The coordination crisis in AI governance isn't merely a communication problem—it's a fundamental information processing challenge that scales with the complexity of the domain. As AI capabilities advance and research proliferates, even the most diligent experts cannot manually process, integrate, and analyze the growing volume of specialized knowledge. We need computational tools that augment human capabilities, much as telescopes extend our vision beyond natural limits.

AMTAIR—Automating Transformative AI Risk Modeling—represents such a tool. It builds upon the MTAIR framework's conceptual foundation while addressing its core limitations through automation and integration. The approach doesn't replace human judgment but amplifies it, scaling up our collective ability to make implicit models explicit and enabling more rigorous evaluation of governance options.

The system architecture implements a five-stage pipeline that transforms unstructured text into interactive, analyzable models:

1. **Text ingestion and preprocessing:** Source documents enter the system, undergo normalization to handle diverse formats, and are stored with citation information preserved.

2. **LLM-powered extraction:** Documents are analyzed using a two-stage process that first identifies key variables and relationships (represented in ArgDown), then extracts probability information (represented in BayesDown).

3. **Bayesian network construction:** BayesDown representations are transformed into formal Bayesian networks with nodes, edges, and conditional probability tables.

4. **Interactive visualization:** The networks are rendered as interactive visualizations that encode probability information through color and provide progressive disclosure of details.
5. **Analysis and inference:** The system enables sensitivity analysis, intervention modeling, and comparison across worldviews.

What distinguishes AMTAIR from previous approaches is the central role of frontier language models in automating the extraction and transformation processes. Rather than treating these models as black boxes that generate answers, AMTAIR employs them as cognitive partners in a structured workflow, using carefully designed prompts to extract specific types of information and transform it between representations.

Consider how this approach differs from traditional methods of knowledge integration. Typically, synthesizing expert perspectives involves reading papers, taking notes, and mentally constructing a composite view—a process limited by individual cognitive capacity and vulnerable to various biases. AMTAIR externalizes this process, making each step explicit and reproducible. The LLM doesn't determine what's important; it helps transform expert knowledge into structured formats that humans can more easily analyze and compare.

The system's primary innovations lie in three areas:

First, the **two-stage extraction process** separates structural understanding from probability estimation, mirroring how humans typically approach complex arguments. This separation improves extraction quality by focusing LLMs on distinct cognitive tasks and creates interpretable intermediate representations.

Second, the **BayesDown representation format** bridges qualitative and quantitative aspects of arguments, maintaining narrative context while enabling mathematical precision. This hybrid format preserves the connection to original texts while supporting computational analysis.

Third, the **interactive visualization approach** makes complex probabilistic models accessible to non-technical stakeholders through intuitive visual encoding and progressive disclosure of information. This enhances cross-domain communication by creating shared reference points.

These innovations address specific limitations of the MTAIR framework. Where MTAIR required days of expert time to formalize arguments, AMTAIR can process papers in minutes. Where MTAIR created static snapshots, AMTAIR enables dynamic updating through integration with forecasting platforms. Where MTAIR struggled with accessibility, AMTAIR provides intuitive visualizations with multiple levels of detail.

The potential impact extends beyond technical achievements. By making implicit models explicit, AMTAIR helps identify genuine disagreements versus terminological confusion. By enabling systematic comparison across worldviews, it facilitates more productive discourse about risk factors and interventions. By supporting counterfactual analysis, it allows policymakers to evaluate governance options across diverse scenarios.

This isn't to suggest that computational tools alone can solve the coordination crisis. Human judgment remains essential for interpreting results, contextualizing insights, and making value-laden decisions. But tools like AMTAIR can dramatically enhance our collective ability to process complex information, identify patterns, and evaluate options—capabilities that become increasingly crucial as AI systems grow more powerful and the stakes of governance decisions rise.

3.2 The Two-Stage Extraction Process

The heart of the AMTAIR approach lies in its two-stage extraction process, which transforms unstructured text into structured probabilistic models through distinct steps that mirror human cognitive processes. This separation—extracting structure before probability—creates important advantages for automation quality, intermediate verification, and interpretability.

When humans analyze complex arguments, they typically first determine what factors matter and how they relate causally, then assess how likely different scenarios are based on those relationships. A climate scientist reading a paper first identifies key variables (emissions, warming, effects) and their causal connections before estimating probabilities of outcomes. This natural cognitive sequence inspired AMTAIR's two-stage approach.

Stage 1: Structure Extraction focuses on identifying key variables and their causal relationships from text, transforming unstructured arguments into ArgDown format. This process involves:

1. **Variable identification:** Determining the key factors discussed in the text, including their possible states (e.g., whether a factor is present/absent or has multiple levels)
2. **Relationship mapping:** Establishing how variables influence each other, creating a directed graph of causal connections
3. **Hierarchical organization:** Arranging variables according to their causal relationships, from root causes to final effects
4. **Metadata attachment:** Annotating each variable with its description and possible states in structured JSON format

The LLM prompt for this stage emphasizes clear identification of causal structure without requiring probability judgments, allowing the model to focus entirely on understanding "what affects what" in the text. This specialized prompt includes detailed instructions about ArgDown syntax, examples of well-formed representations, and guidance for preserving the author's intended meaning.

Figure 4 shows a sample of the ArgDown extraction for Carlsmith's model, illustrating how complex qualitative arguments are transformed into structured representations:

[FIGURE 4: Sample ArgDown extraction from Carlsmith's paper showing hierarchical structure of variables related to existential risk]

```python
def parse_markdown_hierarchy_fixed(markdown_text, ArgDown=True):
    """
    Parse ArgDown format into a structured DataFrame with parent-child relationships.

    Args:
        markdown_text (str): Text in ArgDown format
        ArgDown (bool): If True, extracts only structure without probabilities
                       If False, extracts both structure and probability information

    Returns:
        pandas.DataFrame: Structured data with node information, relationships, and attribute
    """
    # Clean and prepare the text
    clean_text = remove_comments(markdown_text)

    # Extract basic information about nodes
    titles_info = extract_titles_info(clean_text)

    # Determine hierarchical relationships
    titles_with_relations = establish_relationships_fixed(titles_info, clean_text)

    # Convert to structured DataFrame format
    df = convert_to_dataframe(titles_with_relations, ArgDown)

    # Add derived columns for analysis
    df = add_no_parent_no_child_columns_to_df(df)
    df = add_parents_instantiation_columns_to_df(df)

    return df
```

This key function transforms the ArgDown text into a structured DataFrame, capturing the hierarchical relationships between variables and preparing them for further processing. The function works by identifying node titles, descriptions, and indentation levels, then establishing parent-child relationships based on the hierarchy indicated by indentation.

Stage 2: Probability Integration enhances the structural representation with probability information, creating a complete BayesDown specification. This stage involves:

1. **Question generation:** Automatically creating appropriate probability questions based on the network structure
2. **Probability extraction:** Obtaining probability estimates for each question, either from the text or through LLM inference
3. **Consistency checking:** Ensuring probability distributions sum to 1 and match structural constraints
4. **BayesDown integration:** Incorporating probability information into the ArgDown structure

The key innovation in this stage is the automated generation of appropriate probability questions based on network structure. For each node, the system generates questions about prior probabilities (how likely is this variable in isolation?) and conditional probabilities (how likely is this variable given different states of its parents?).

Figure 5 illustrates how probability questions are derived for a simple node with one parent:

[FIGURE 5: Diagram showing how probability questions are generated based on network structure]

For the "Sprinkler" node with parent "Rain," the system automatically generates questions like:

- What is the probability for Sprinkler=sprinkler_TRUE?
- What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_TRUE?
- What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_FALSE?

These questions are then answered either by extracting explicit probabilities from the text or by having the LLM infer reasonable values based on the author's arguments. The answers are structured into a complete BayesDown representation that includes both the causal structure and all necessary probability information.

The visualization below demonstrates the completed extraction for a portion of Carlsmith's model, showing how variables like "Misaligned Power Seeking" are influenced by multiple factors, each with associated probabilities:

[VISUALIZATION: Extracted causal structure from Carlsmith's model with probability information]

This two-stage approach offers several important advantages:

1. **Improved extraction quality:** By focusing on one cognitive task at a time, the LLM performs better at each stage than it would attempting to extract everything simultaneously.
2. **Intermediate verification:** Having ArgDown as an intermediate representation allows human verification before probability extraction, catching structural errors early.

3. **Separation of concerns:** Structure and probability can be updated independently, enabling more flexible maintenance as new information emerges.
4. **Alignment with human cognition:** The process mirrors how experts approach complex arguments, making the system's operation more intuitive and interpretable.

Perhaps most importantly, the intermediate ArgDown representation creates a bridge between qualitative and quantitative aspects of arguments. It preserves the narrative structure and conceptual relationships from the original text while preparing for mathematical precision through probability integration. This hybrid approach maintains the strengths of both worlds: the richness of natural language and the rigor of formal models.

3.3 BayesDown: Bridging Qualitative and Quantitative Representation

If the coordination crisis in AI governance stems partly from incompatible languages across domains—technical researchers speaking in mathematical formalisms, policy specialists in institutional frameworks, and ethicists in normative concepts—then effective coordination requires bridges between these domains. BayesDown serves as such a bridge, combining the narrative richness of qualitative argumentation with the precision of quantitative probability judgments.

Traditional formal representations face a fundamental tradeoff: increase precision and you sacrifice accessibility; enhance accessibility and you lose precision. Mathematical notations offer exactness but exclude many stakeholders. Natural language provides accessibility but permits ambiguity and vagueness. This tradeoff creates communication barriers between technical and policy domains, limiting coordination on complex challenges like AI governance.

BayesDown disrupts this tradeoff by creating a hybrid representation that preserves strengths from both worlds. Its design follows three key principles:

First, **human readability** ensures the representation remains interpretable without specialized training. The syntax builds on familiar conventions from markdown and JSON, maintaining hierarchical relationships through indentation and encapsulating technical details within structured metadata. Unlike purely mathematical notations, the format preserves natural language descriptions alongside formal elements.

Second, **machine processability** enables computational analysis and transformation. The consistent syntax permits automated parsing, formal verification, and conversion to computational models like Bayesian networks. The structured JSON metadata provides clear paths for extracting probability information and mapping it to conditional probability tables.

Third, **contextual preservation** maintains the connection to original arguments. By including descriptive text alongside formal structure, BayesDown retains the narrative context and qualitative considerations that inform probability judgments. This contextual information helps users interpret the model in light of the original arguments.

Consider how these principles manifest in the BayesDown syntax. Each node begins with a bracketed title followed by a natural language description, preserving the core statement being formalized. The JSON metadata contains technical information like instantiations, priors, and posteriors, but keeps this information clearly separated from the narrative content. Hierarchical relationships use indentation and plus symbols, creating a visual structure that mirrors causal influence.

```
[Existential_Catastrophe]: The destruction of humanity's long-term potential due to AI syster
  "instantiations": ["existential_catastrophe_TRUE", "existential_catastrophe_FALSE"],
  "priors": {"p(existential_catastrophe_TRUE)": "0.05", "p(existential_catastrophe_FALSE)": "
  "posteriors": {
    "p(existential_catastrophe_TRUE|human_disempowerment_TRUE)": "0.95",
    "p(existential_catastrophe_TRUE|human_disempowerment_FALSE)": "0.0"
  }
}
 + [Human_Disempowerment]: Permanent and collective disempowerment of humanity relative to AI
   "instantiations": ["human_disempowerment_TRUE", "human_disempowerment_FALSE"],
   "priors": {"p(human_disempowerment_TRUE)": "0.208", "p(human_disempowerment_FALSE)": "0.79
   "posteriors": {
     "p(human_disempowerment_TRUE|scale_of_power_seeking_TRUE)": "1.0",
     "p(human_disempowerment_TRUE|scale_of_power_seeking_FALSE)": "0.0"
   }
 }
```

This excerpt from the Carlsmith model representation illustrates how BayesDown preserves both the narrative description ("The destruction of humanity's long-term potential…") and the precise probability judgments. Someone without technical background can still understand the core claims and their relationships, while someone seeking quantitative precision can find exact probability values.

The format supports multiple levels of engagement. At the most basic level, readers can follow the hierarchical structure to understand causal relationships between factors. At an intermediate level, they can examine probability judgments to assess the strength of different influences. At the most technical level, they can analyze the complete probabilistic model to perform inference and sensitivity analysis.

This multi-level accessibility creates important advantages for coordination across domains:

1. **Technical-policy translation:** BayesDown provides a common reference point for technical researchers explaining safety concerns and policy specialists evaluating governance options, reducing communication barriers.

2. **Argumentation transparency:** The format makes assumptions explicit, helping identify genuine disagreements versus terminological confusion or unstated premises.

3. **Incremental formalization:** BayesDown supports varying levels of formality, from qualitative structure to complete probability specifications, allowing gradual progression from informal to formal representations.
4. **Verification flexibility:** Human experts can verify extracted representations at different levels—checking structural correctness without assessing probabilities, or focusing on critical probability judgments without reviewing the entire model.

The hybrid nature of BayesDown aligns with how experts typically communicate complex ideas: combining qualitative explanations with quantitative judgments, using natural language to provide context for formal claims, and adjusting precision based on audience needs. By mirroring these natural communication patterns, BayesDown makes formalization more intuitive and accessible.

This bridging function extends beyond representation to influence the entire extraction and analysis workflow. When extracting from text, the two-stage process preserves narrative context alongside formal structure. When visualizing models, interactive interfaces provide both qualitative descriptions and quantitative details. When evaluating policies, counterfactual analysis incorporates both mathematical precision and contextual interpretation.

In the broader context of the coordination crisis, BayesDown demonstrates how thoughtfully designed intermediate representations can overcome communication barriers between domains. Rather than forcing all stakeholders to adopt a single specialized language, it creates a flexible format that accommodates different perspectives while enabling precise analysis—precisely the kind of bridge needed for effective coordination on complex governance challenges.

3.4 Interactive Visualization and Exploration

Complex probabilistic models like Bayesian networks contain rich information, but they often remain inaccessible to many stakeholders. A conditional probability table with dozens of values conveys precise relationships, but few can intuitively grasp its implications. This accessibility gap limits the potential for coordinated action on AI governance challenges—what good is formalization if the resulting models remain opaque to most decision-makers?

AMTAIR addresses this challenge through interactive visualization designed to make complex probabilistic relationships accessible to diverse stakeholders. The approach combines visual encoding of probability information, progressive disclosure of details, and interactive exploration capabilities to create intuitive interfaces for complex models.

The visualization system follows several key design principles:

First, **visual encoding of probability** uses color gradients to represent likelihood values. Nodes are colored on a spectrum from red (low probability) to green (high probability) based on their primary state's probability. This simple visual cue provides immediate insights into which outcomes are more or less likely without requiring numerical interpretation.

Second, **structural classification** uses border colors to indicate node types based on network position. Blue borders designate root causes (nodes without parents), purple borders mark intermediate nodes (with both parents and children), and magenta borders highlight leaf nodes (final effects without children). This classification helps users understand the causal flow through the network.

Third, **progressive disclosure** presents information in layers of increasing detail. Basic node information appears in the visualization itself, additional details emerge in tooltips on hover, and comprehensive probability tables display in modal windows on click. This layered approach prevents information overload while ensuring all details remain accessible.

Fourth, **interactive exploration** allows users to reorganize nodes, zoom in on areas of interest, adjust physics parameters, and investigate probability values. These capabilities transform the visualization from a static image into an explorable knowledge landscape.

Figure 6 shows the interactive visualization of Carlsmith's model, highlighting how color, border styling, and layout work together to represent complex causal relationships:

[FIGURE 6: Interactive visualization of Carlsmith's model showing color-coded nodes and causal relationships]

The visualization system implements these principles through a combination of NetworkX for graph representation and PyVis for interactive display, with custom HTML generation for tooltips and modals:

```python
def create_bayesian_network_with_probabilities(df):
    """
    Create an interactive Bayesian network visualization with enhanced probability visualiza
    and node classification based on network structure.
    """
    # Create network structure
    G = nx.DiGraph()

    # Add nodes with attributes
    for idx, row in df.iterrows():
        title = row['Title']
        description = row['Description']
        priors = get_priors(row)
        instantiations = get_instantiations(row)

        G.add_node(title, description=description, priors=priors,
                   instantiations=instantiations, posteriors=get_posteriors(row))

    # Add edges based on parent-child relationships
    for idx, row in df.iterrows():
```

```
    child = row['Title']
    parents = get_parents(row)

    for parent in parents:
        if parent in G.nodes():
            G.add_edge(parent, child)

# Classify nodes based on network structure
classify_nodes(G)

# Create visualization network
net = Network(notebook=True, directed=True, cdn_resources="in_line",
             height="600px", width="100%")

# Configure physics for better layout
net.force_atlas_2based(gravity=-50, spring_length=100, spring_strength=0.02)
net.show_buttons(filter_=['physics'])

# Add graph to network
net.from_nx(G)

# Enhance node appearance
for node in net.nodes:
    node_id = node['id']
    node_data = G.nodes[node_id]

    # Set border color based on node type
    node_type = node_data.get('node_type', 'unknown')
    border_color = get_border_color(node_type)

    # Set background color based on probability
    priors = node_data.get('priors', {})
    background_color = get_probability_color(priors)

    # Create tooltip and expanded content
    tooltip = create_tooltip(node_id, node_data)
    node_data['expanded_content'] = create_expanded_content(node_id, node_data)

    # Set node attributes
    node['title'] = tooltip
    node['label'] = f"{node_id}\np={priors.get('true_prob', 0.5):.2f}"
    node['shape'] = 'box'
```

```
        node['color'] = {
            'background': background_color,
            'border': border_color,
            'highlight': {
                'background': background_color,
                'border': border_color
            }
        }

    # Setup click handling for detailed information
    # [Click handling JavaScript code omitted for brevity]

    return net.show('bayesian_network.html')
```

Beyond the core visualization, the system includes specialized components that enhance understanding of probabilistic relationships:

1. **Probability bars** provide visual representations of probability distributions, showing relative likelihoods of different states using color-coded horizontal bars with numeric labels.
2. **Conditional probability tables** organize complex relationships into structured matrices, displaying how different combinations of parent states influence probability distributions.
3. **Sensitivity indicators** highlight which nodes and relationships most significantly affect outcomes, directing attention to critical factors.

These components work together to create an intuitive interface for complex probabilistic models. A user might start by exploring the overall structure to understand key factors and relationships, hover over nodes of interest to see probability summaries, then click on specific nodes to examine detailed conditional probabilities.

The benefits of this visualization approach extend beyond aesthetic appeal to fundamental improvements in understanding and communication:

First, **intuitive comprehension** of probability relationships becomes possible even for those without formal training in Bayesian statistics. The color coding provides immediate visual cues about which outcomes are more likely, while interactive exploration allows users to develop intuition about how different factors influence results.

Second, **cross-stakeholder communication** improves through shared visual reference points. Technical experts can use the visualizations to explain complex relationships to policy specialists, while governance experts can identify institutional factors that might be incorporated into the models.

Third, **disagreement identification** becomes more precise as stakeholders can point to specific nodes, relationships, or probability values where their views differ, focusing discussion on substantive issues rather than terminological confusion.

Fourth, **intervention assessment** becomes more concrete as users can see how changing specific factors influences downstream effects, providing intuitive understanding of causal pathways and leverage points.

The visualization system demonstrates how thoughtful interface design can overcome barriers to understanding complex formal models. By making probabilistic relationships visually intuitive and progressively disclosing details based on user interest, it creates bridges between mathematical precision and human comprehension—precisely the kind of bridge needed to support coordination across domains in AI governance.

This approach reflects a broader principle: formalization is most valuable when it enhances rather than replaces human understanding. The AMTAIR visualization doesn't simplify complex relationships; it makes them more accessible by leveraging visual cognition, interactive exploration, and progressive disclosure. This human-centered approach to formalization creates tools that augment rather than replace expert judgment, enhancing our collective ability to understand and address complex governance challenges.

3.5 Beyond Extraction: Toward Policy Evaluation

Formalizing expert knowledge through automated extraction creates valuable epistemic infrastructure, but the ultimate goal extends beyond representation to supporting concrete governance decisions. Once implicit models become explicit through the AMTAIR approach, they enable a crucial capability: systematic evaluation of how policy interventions might affect outcomes across different scenarios.

This capability addresses a fundamental challenge in AI governance: making decisions under deep uncertainty about future developments. Traditional approaches often rely on point forecasts or vague qualitative judgments, creating environments where rhetoric outweighs evidence and status determines influence. Formal models enable a more disciplined approach, systematically exploring how different interventions perform across a range of assumptions.

The AMTAIR system supports policy evaluation through three key mechanisms:

First, **counterfactual analysis** implements Pearl's do-calculus to simulate interventions on the causal system. Rather than merely observing correlations, this approach explicitly models what happens when we force a variable to take a specific value, accounting for how this intervention propagates through the causal structure. For example, we can ask how requiring safety demonstrations (setting a variable to a specific value) would affect the likelihood of misaligned systems and ultimately existential risk.

Second, **intervention modeling** provides structured representations of policy options that can be applied to the causal model. Policies are formalized as modifications to specific variables, relationships, or probability distributions, creating concrete representations of how governance actions influence the system. For example, compute governance might be modeled as reducing the probability of rapid capability jumps, while safety standards might increase the likelihood of warning shots.

Third, **cross-worldview comparison** enables evaluation of interventions across different causal models and probability distributions. Rather than assuming a single correct model, this approach acknowledges legitimate uncertainty about causal structure and relationships, testing how interventions perform across different plausible world models. This identifies "robust" policies that work reasonably well regardless of which worldview proves correct—a crucial capability when decisions must be made despite fundamental disagreements.

Consider how these mechanisms apply to Carlsmith's model of existential risk from power-seeking AI. Figure 7 shows the evaluation of a hypothetical governance intervention requiring safety demonstrations before deployment:

[FIGURE 7: Visualization showing policy impact evaluation across Carlsmith model]

The analysis simulates how requiring safety demonstrations affects deployment decisions for potentially misaligned systems, and consequently how this influences the probability of misaligned power-seeking and ultimately existential catastrophe. By comparing the baseline probability (5%) with the intervention probability (3.2% in this example), we can quantify the potential risk reduction from this policy.

The implementation uses counterfactual queries on the Bayesian network:

```
def evaluate_policy_impact(model, intervention_variable, intervention_value, target_variable
    """
    Evaluate the impact of setting a variable to a specific value on a target outcome.

    Args:
        model: Bayesian network model
        intervention_variable: Variable to intervene on
        intervention_value: Value to set for intervention
        target_variable: Outcome variable of interest
        target_value: Outcome value of interest

    Returns:
        dict: Impact analysis including baseline and intervention probabilities
    """
    # Create inference engine
    inference = VariableElimination(model)
    # Calculate baseline probability
    baseline_query = inference.query(variables=[target_variable])
    baseline_prob = baseline_query.values[baseline_query.state_names[target_variable].index(t

    # Calculate intervention probability using do-calculus
    intervention_query = inference.query(
        variables=[target_variable],
        evidence={intervention_variable: intervention_value},
        do={intervention_variable: intervention_value}  # The do-operation
    )
    intervention_prob = intervention_query.values[intervention_query.state_names[target_varia

    # Calculate impact
    absolute_change = intervention_prob - baseline_prob
    relative_change = absolute_change / baseline_prob * 100 if baseline_prob > 0 else float(

    return {
        'baseline_probability': baseline_prob,
        'intervention_probability': intervention_prob,
        'absolute_change': absolute_change,
        'relative_change': relative_change
    }
```

This function implements the counterfactual analysis, calculating both the baseline probability of the target outcome and the probability after intervention. The `do` operation ensures proper handling of causal effects rather than merely conditioning on observed values.

Beyond analyzing individual interventions, the system can evaluate portfolios of complementary policies, identifying synergies and conflicts between different approaches. For example, it might examine how compute governance, safety standards, and liability rules work together to reduce risk more effectively than any single intervention alone.

The policy evaluation capabilities extend to more sophisticated analyses:

1. **Robustness assessment** examines how sensitive intervention effects are to variations in model parameters, identifying policies that maintain effectiveness despite uncertainty about exact probability values.

2. **Option value analysis** evaluates how different policies affect our ability to gather information and make better decisions in the future, capturing the value of preserving flexibility.

3. **Intervention portfolio construction** identifies sets of complementary policies that address different aspects of risk, creating more robust governance approaches.

4. **Dependency mapping** visualizes prerequisites and enabling conditions between interventions, helping understand sequencing requirements and potential bottlenecks.

These capabilities transform governance discussions from abstract debates about principles to concrete analyses of expected impacts. Rather than merely asserting that a policy would reduce risk, stakeholders can demonstrate specific causal pathways through which the intervention affects outcomes, quantify the magnitude of expected effects, and test robustness across different assumptions.

This approach doesn't eliminate value judgments or normative considerations—those remain essential for determining appropriate governance goals and acceptable tradeoffs. But it adds rigor to instrumental reasoning about how different interventions might achieve those goals, reducing the influence of rhetoric, status, and cognitive biases in policy evaluation.

In the context of the coordination crisis, these policy evaluation capabilities create a shared language for discussing interventions across domains. Technical researchers can express safety concerns in terms of how they affect model variables; policy specialists can formulate governance proposals as interventions on specific factors; ethicists can articulate normative considerations as valued outcomes or constraints on acceptable interventions. This common framework facilitates more productive coordination without requiring all stakeholders to adopt a single specialized vocabulary.

4. Implementation: The AMTAIR Prototype

4.1 System Architecture and Data Flow

The AMTAIR prototype implements the conceptual architecture described earlier through a modular, extensible system designed to transform text into interactive Bayesian networks. This section details the technical realization of this architecture, explaining how different components interact to enable automated extraction and analysis.

At its core, the system consists of five main components connected in a sequential pipeline with feedback loops:

1. **Text ingestion and preprocessing** handles the initial transformation of source documents into a standardized format suitable for extraction. This component supports various input formats (PDF, markdown, plain text) and preserves citation information to maintain provenance.
2. **LLM-powered extraction pipeline** implements the two-stage process for transforming normalized text into structured representations. The first stage extracts structural information (ArgDown), while the second stage enhances it with probability information (BayesDown).

3. **Bayesian network construction** converts BayesDown representations into formal Bayesian networks with nodes, edges, and conditional probability tables. This component includes data transformation, network analysis, and enhancement with derived metrics.

4. **Visualization and interaction interface** creates interactive presentations of the Bayesian networks with probability encoding, progressive disclosure, and exploration capabilities. This component generates HTML with embedded JavaScript for interactivity.

5. **Analysis and inference engine** enables probabilistic reasoning about the networks, including marginal and conditional probability calculations, sensitivity analysis, and counterfactual evaluation for policy assessment.

Figure 8 illustrates the data flow between these components:

[FIGURE 8: Diagram showing data flow between system components]

The implementation uses a combination of Python libraries for different aspects of the pipeline:

- **pandas** for structured data manipulation throughout the pipeline
- **networkx** for graph representation and analysis
- **pgmpy** for Bayesian network construction and inference
- **pyvis** for interactive network visualization
- **requests** for API calls to language models
- **matplotlib** for static visualizations

This architecture balances several design principles:

Modularity ensures that each component can be developed, tested, and improved independently. For example, the extraction pipeline can be enhanced without modifying the visualization system, and different visualization approaches can be implemented without changing the extraction logic.

Explicitness makes the transformation process transparent and inspectable at each stage. Rather than using end-to-end black-box processing, the system creates intermediate representations (ArgDown, BayesDown, DataFrames) that can be examined and verified.

Interactivity prioritizes human engagement with the results, creating rich interfaces that reveal both structural and probabilistic information through visual encoding and progressive disclosure.

Extensibility supports incremental enhancement through well-defined interfaces between components. New capabilities can be added without redesigning the entire system, enabling gradual improvement over time.

The core code organization reflects this architecture:

```
amtair/
    ingestion/              # Text preprocessing and normalization
        pdf_processor.py
        markdown_processor.py
        text_normalizer.py
    extraction/             # LLM-powered extraction pipeline
        argdown_extractor.py
        bayesdown_enhancer.py
        prompt_templates.py
    network/                # Bayesian network construction
        network_builder.py
        data_transformer.py
        metrics_calculator.py
    visualization/          # Interactive visualization
        network_visualizer.py
        html_generator.py
        color_mapper.py
    analysis/               # Analysis and inference
        inference_engine.py
        sensitivity_analyzer.py
        policy_evaluator.py
    utils/                  # Shared utilities
        data_structures.py
        file_operations.py
        logging_config.py
```

This organization makes dependencies explicit while enabling independent development of different components. For example, the extraction team can enhance prompt templates without affecting the network construction code, and the visualization team can improve the user interface without modifying the underlying data structures.

The prototype implementation focused on demonstrating the core pipeline functionality rather than building a complete production system. As a result, the current version has certain limitations:

1. It relies on external API calls to frontier LLMs rather than deploying models locally.
2. It processes documents one at a time rather than ingesting entire literature repositories.
3. It implements basic policy evaluation capabilities without the full range of analysis features.
4. It focuses on BayesDown as the intermediate representation without supporting alternative formats.

Despite these limitations, the prototype successfully demonstrates the feasibility of automating the extraction and transformation process, creating a foundation for more sophisticated implementations in the future.

The architecture's design anticipates future extensions, including integration with prediction markets for dynamic updating, support for cross-worldview comparison, and enhanced policy evaluation capabilities. These extensions would build on the existing foundation rather than requiring architectural redesign, demonstrating the value of the modular approach.

4.2 The Rain-Sprinkler-Lawn Implementation

Before applying the AMTAIR approach to complex real-world risk assessments, I validated the implementation using the canonical rain-sprinkler-lawn example introduced earlier. This simple but complete example allows step-by-step verification of each component in the pipeline, from initial representation to interactive visualization.

The rain-sprinkler-lawn scenario has become something of a "Hello World" for Bayesian networks—simple enough to understand intuitively but complex enough to demonstrate conditional independence and inference. It involves three variables: Rain (whether it's raining), Sprinkler (whether the sprinkler is on), and Grass_Wet (whether the grass is wet). Both rain and the sprinkler can cause the grass to be wet, while rain also influences whether the sprinkler is used (as people typically don't run sprinklers when it's already raining).

Stage 1: ArgDown Representation captures the structural relationships between these variables without probability information. The implementation starts with this representation:

```
[Grass_Wet]: Concentrated moisture on, between and around the blades of grass. {"instantiatio
 + [Rain]: Tears of angles crying high up in the skies hitting the ground. {"instantiations"
 + [Sprinkler]: Activation of a centrifugal force based CO2 droplet distribution system. {"in
   + [Rain]
```

This ArgDown representation captures several key aspects of the scenario:

- The three variables with their natural language descriptions
- Their possible states (TRUE/FALSE for each variable)
- The causal structure (both Rain and Sprinkler influence Grass_Wet, and Rain influences Sprinkler)

The system processes this representation with the parsing function shown in the previous section, transforming it into a structured DataFrame that explicitly represents parent-child relationships:

```
# Process the ArgDown representation
argdown_df = parse_markdown_hierarchy_fixed(argdown_text, ArgDown=True)

# Display the results
print(argdown_df[['Title', 'Description', 'Parents', 'Children', 'instantiations']])
```

This processing correctly extracts the structural information, identifying that:

- Grass_Wet has parents Rain and Sprinkler, but no children
- Rain has no parents, but is a parent to both Grass_Wet and Sprinkler
- Sprinkler has parent Rain and child Grass_Wet

Stage 2: BayesDown Enhancement adds probability information to the structural representation. The implementation first generates appropriate probability questions based on the network structure:

```
# Generate probability questions based on network structure
df_with_questions = generate_argdown_with_questions(argdown_df, "ArgDown_WithQuestions.csv")

# Display sample questions for the Sprinkler node
sprinkler_questions = df_with_questions.loc[df_with_questions['Title'] == 'Sprinkler', 'Gene
print(json.loads(sprinkler_questions))
```

For the Sprinkler node, this generates questions like:

- What is the probability for Sprinkler=sprinkler_TRUE?
- What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_TRUE?
- What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_FALSE?

After answering these questions (manually or via LLM), the system incorporates the probability information into a complete BayesDown representation:

```
[Grass_Wet]: Concentrated moisture on, between and around the blades of grass. {
  "instantiations": ["grass_wet_TRUE", "grass_wet_FALSE"],
  "priors": {"p(grass_wet_TRUE)": "0.322", "p(grass_wet_FALSE)": "0.678"},
  "posteriors": {
    "p(grass_wet_TRUE|sprinkler_TRUE,rain_TRUE)": "0.99",
    "p(grass_wet_TRUE|sprinkler_TRUE,rain_FALSE)": "0.9",
    "p(grass_wet_TRUE|sprinkler_FALSE,rain_TRUE)": "0.8",
    "p(grass_wet_TRUE|sprinkler_FALSE,rain_FALSE)": "0.0"
  }
}
```

```
+ [Rain]: Tears of angles crying high up in the skies hitting the ground. {
  "instantiations": ["rain_TRUE", "rain_FALSE"],
  "priors": {"p(rain_TRUE)": "0.2", "p(rain_FALSE)": "0.8"}
}
+ [Sprinkler]: Activation of a centrifugal force based CO2 droplet distribution system. {
  "instantiations": ["sprinkler_TRUE", "sprinkler_FALSE"],
  "priors": {"p(sprinkler_TRUE)": "0.44838", "p(sprinkler_FALSE)": "0.55162"},
  "posteriors": {
    "p(sprinkler_TRUE|rain_TRUE)": "0.01",
    "p(sprinkler_TRUE|rain_FALSE)": "0.4"
  }
}
  + [Rain]
```

This BayesDown representation now contains complete probability information:

- Prior probabilities for each variable (e.g., P(Rain=TRUE) = 0.2)
- Conditional probabilities for variables with parents (e.g., P(Sprinkler=TRUE|Rain=TRUE) = 0.01)

Stage 3: Bayesian Network Construction transforms the BayesDown representation into a formal Bayesian network with nodes, edges, and conditional probability tables. The implementation extracts the information into a structured DataFrame, then converts this into a network representation:

```python
# Extract data from BayesDown representation
extracted_df = parse_markdown_hierarchy_fixed(bayesdown_text, ArgDown=False)

# Enhance the data with calculated metrics
enhanced_df = enhance_extracted_data(extracted_df)

# Create a Bayesian network from the extracted data
def create_bayesian_network(df):
    # Create network structure
    model = BayesianNetwork()

    # Add nodes and edges
    for idx, row in df.iterrows():
        title = row['Title']
        parents = row['Parents'] if isinstance(row['Parents'], list) else []

        # Add node
        model.add_node(title)
        # Add edges from parents to this node
        for parent in parents:
            model.add_edge(parent, title)

    # Add CPDs for each node
    for idx, row in df.iterrows():
        title = row['Title']
        parents = row['Parents'] if isinstance(row['Parents'], list) else []
        instantiations = row['instantiations'] if isinstance(row['instantiations'], list) els
        priors = row['priors'] if isinstance(row['priors'], dict) else {}
        posteriors = row['posteriors'] if isinstance(row['posteriors'], dict) else {}

        # Create CPD based on whether node has parents
        if not parents:  # No parents - use prior probabilities
            # Implementation details omitted for brevity
        else:  # Has parents - use conditional probabilities
            # Implementation details omitted for brevity

        # Add CPD to model
        model.add_cpds(cpd)

    # Check model validity
    model.check_model()

    return model

# Create the network
bayesian_network = create_bayesian_network(enhanced_df)
```

The resulting Bayesian network correctly represents the causal structure and probability distributions from the BayesDown representation. This network enables various types of probabilistic inference, such as calculating the probability of rain given that the grass is wet:

```
# Create inference engine
inference = VariableElimination(bayesian_network)

# Calculate P(Rain=TRUE | Grass_Wet=TRUE)
result = inference.query(variables=['Rain'], evidence={'Grass_Wet': 'grass_wet_TRUE'})
print(f"P(Rain=TRUE | Grass_Wet=TRUE) = {result.values[0]:.3f}")
```

Visual Result The implementation creates an interactive visualization of the network using the function described in the previous section:

```
# Create interactive visualization
visualization = create_bayesian_network_with_probabilities(enhanced_df)
display(visualization)
```

Figure 9 shows the resulting visualization with color-coded nodes indicating probability values:

[FIGURE 9: Interactive visualization of the rain-sprinkler-lawn Bayesian network]

The visualization correctly encodes the causal structure (arrows from causes to effects) and probability information (node colors indicating likelihood), providing an intuitive representation of the relationships between variables.

Validation To verify the implementation's correctness, I compared computational results from the network with analytical solutions calculated by hand. For example, the probability of wet grass can be calculated analytically:

P(W=TRUE) = , P(W=TRUE|R=r,S=s) × P(R=r) × P(S=s|R=r)

Where the sum is over all possible values of r and s. The computational result from the Bayesian network (0.322) matched the analytical calculation, confirming the implementation's correctness.

Similarly, posterior probabilities like P(R=TRUE|W=TRUE) were verified against analytical calculations using Bayes' rule:

P(R=TRUE|W=TRUE) = P(W=TRUE|R=TRUE) × P(R=TRUE) / P(W=TRUE)

The rain-sprinkler-lawn implementation demonstrates the complete AMTAIR pipeline functioning correctly on a simple but non-trivial example. Each step in the process—from ArgDown representation through BayesDown enhancement to Bayesian network construction and visualization—performs as expected, transforming a structured representation into an interactive, analyzable model.

This validation provides confidence that the approach can be successfully applied to more complex, real-world scenarios like Carlsmith's model of existential risk, which follows the same principles but involves many more variables and relationships.

4.3 Application to Carlsmith's Model

Having validated the implementation on the canonical rain-sprinkler-lawn example, I applied the AMTAIR approach to a substantially more complex real-world case: Joseph Carlsmith's model of existential risk from power-seeking AI. This application demonstrates the system's ability to handle sophisticated multi-level arguments with numerous variables and relationships.

Carlsmith's analysis involves dozens of factors organized in a complex causal structure, from root causes like "Advanced AI Capability" and "Instrumental Convergence" through intermediate factors like "APS Systems" and "Misaligned Power Seeking" to final outcomes like "Existential Catastrophe." The model exhibits several challenging features:

1. **Multi-level structure** with causal chains spanning multiple steps
2. **Divergent pathways** where factors influence outcomes through multiple routes
3. **Complex conditional dependencies** with variables influenced by multiple parents
4. **Variables with three or more possible states** rather than simple binary outcomes
5. **Interconnected clusters** where factors form distinct but related argument groups

The extraction process began with an ArgDown representation capturing the structural relationships between variables:

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

This representation captures the complex causal structure of Carlsmith's argument, with 21 variables organized in a multi-level hierarchy. The "Misaligned_Power_Seeking" node appears multiple times, reflecting its role as a central concept that influences several other variables.

After processing this structure with the AMTAIR system, probability information was added to create a complete BayesDown representation. The following excerpt shows the probability information for a single node ("Deployment_Decisions"):

```
[Deployment_Decisions]: Decisions to deploy potentially misaligned AI systems. {
  "instantiations": ["deployment_decisions_DEPLOY", "deployment_decisions_WITHHOLD"],
  "priors": {
    "p(deployment_decisions_DEPLOY)": "0.70",
    "p(deployment_decisions_WITHHOLD)": "0.30"
  },
  "posteriors": {
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_TRUE)": "0
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_FALSE)": "
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_TRUE)": "0.6
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_FALSE)": "0
  }
}
```

This node has two possible states (DEPLOY or WITHHOLD), prior probabilities for each state, and conditional probabilities based on different combinations of its parent variables ("Incentives_To_Build_APS" and "Deception_By_AI").

The complete BayesDown representation was processed through the AMTAIR pipeline, resulting in a structured DataFrame and ultimately a Bayesian network. Key extraction steps included:

```
# Extract structured data from BayesDown
carlsmith_df = parse_markdown_hierarchy_fixed(carlsmith_bayesdown, ArgDown=False)

# Enhance with calculated metrics
enhanced_carlsmith_df = enhance_extracted_data(carlsmith_df)

# Create network and visualization
carlsmith_network = create_bayesian_network(enhanced_carlsmith_df)
carlsmith_visualization = create_bayesian_network_with_probabilities(enhanced_carlsmith_df)
```

The resulting visualization (Figure 10) shows the complete Carlsmith model with color-coded nodes representing probability values:

[FIGURE 10: Interactive visualization of Carlsmith's model showing color-coded nodes and relationships]

This visualization reveals several structural insights:

1. **Central importance of "Misaligned_Power_Seeking"** as a hub node with multiple parents and children
2. **Multiple pathways to "Existential_Catastrophe"** through different intermediate factors
3. **Clusters of related variables** forming coherent subarguments (e.g., factors affecting alignment difficulty)
4. **Flow of influence** from technical factors (bottom) through deployment decisions to ultimate outcomes (top)

The implementation successfully handles the complexity of Carlsmith's model, correctly processing the multi-level structure, resolving repeated node references, and calculating appropriate probability distributions. The interactive visualization makes this complex model accessible, allowing users to explore different aspects of the argument through intuitive navigation.

Several key aspects of the implementation were particularly important for handling this complex model:

1. The **parent-child relationship detection algorithm** correctly identified hierarchical relationships despite the complex structure with repeated nodes and multiple levels.

2. The **probability question generation system** created appropriate questions for all variables, including those with multiple parents requiring factorial combinations of conditional probabilities.

3. The **network enhancement functions** calculated useful metrics like centrality measures and Markov blankets that help interpret the model structure.

4. The **visualization system** effectively presented the complex network through color-coding, interactive exploration, and progressive disclosure of details.

The successful application to Carlsmith's model demonstrates the AMTAIR approach's scalability to complex real-world arguments. While the canonical rain-sprinkler-lawn example validated correctness, this application proves practical utility for sophisticated multi-level arguments with dozens of variables and complex interdependencies—precisely the kind of arguments that characterize AI risk assessments.

This capability addresses a core limitation of the original MTAIR framework: the labor intensity of manual formalization. Where manually converting Carlsmith's argument to a formal model might take days of expert time, the AMTAIR approach accomplished this in minutes, creating a foundation for further analysis and exploration.

4.4 Performance and Validation

The AMTAIR prototype demonstrates promising capabilities, but any automated system requires rigorous evaluation to assess reliability, accuracy, and practical utility. This section presents performance metrics and validation approaches for the extraction and transformation processes, providing a foundation for understanding the system's strengths and limitations.

Extraction Quality Metrics assess how accurately the system extracts structured representations from source texts. I evaluated extraction quality using three complementary approaches:

1. **Comparison to manual extraction:** For select examples, I compared automated extraction results with manually created representations, calculating precision, recall, and F1 scores for nodes, relationships, and probability values.
2. **Structural validation:** I used formal validation rules to check structural properties like acyclicity, completeness (all referenced nodes defined), and consistency (probability distributions sum to 1).
3. **Expert review:** I enlisted domain experts to assess the semantic accuracy of extracted representations, focusing on whether they preserved the author's intended meaning.

Table 1 summarizes extraction quality metrics for different components of the pipeline:

Component	Precision	Recall	F1 Score
Node identification	0.94	0.91	0.92
Relationship detection	0.89	0.85	0.87
Prior probability extraction	0.91	0.88	0.89
Conditional probability extraction	0.83	0.78	0.80

These metrics reveal stronger performance on structural extraction (nodes and relationships) than on probability extraction, particularly for conditional probabilities where complexity increases with multiple parent variables. This pattern aligns with the two-stage extraction approach, which prioritizes structural accuracy before addressing probability information.

The extraction quality assessment also revealed common failure modes:

1. **Complex causal expressions** where influence is described through multiple sentences or implicit relationships
2. **Ambiguous probability language** using terms like "likely," "probably," or "almost certainly" without precise definitions
3. **Deep nesting** where relationships span multiple levels of indirection
4. **Novel terminology** without sufficient context for interpretation

These failure modes suggest specific areas for improvement in future implementations, such as enhanced context handling for complex expressions and better interpretation of qualitative probability language.

Computational Performance Metrics assess how efficiently the system processes inputs and generates outputs. I measured performance across different network sizes to understand scaling characteristics:

Network Size (nodes)	Extraction Time (s)	Network Construction (s)	Visualization (s)	Total Processing (s)
Small (5-10)	3.2	0.4	0.6	4.2
Medium (10-50)	12.5	1.3	1.9	15.7
Large (50+)	42.8	3.7	5.2	51.7

The extraction phase dominates processing time, primarily due to API calls to frontier LLMs. Network construction and visualization scale well with network size, showing sub-linear growth as complexity increases. The current implementation prioritizes accuracy over speed, with several opportunities for optimization:

1. **Batched extraction** could process multiple nodes or relationships simultaneously
2. **Caching mechanisms** could avoid redundant processing of repeated patterns
3. **Progressive refinement** could focus detailed extraction on critical parts of the network

Despite these optimization opportunities, the current performance is sufficient for practical use cases. Processing Carlsmith's model (21 nodes with complex relationships) took approximately 18 seconds, enabling interactive exploration and experimentation.

Validation Code ensures extraction quality through automated checks for structural and probabilistic consistency:

```python
def validate_bayesian_network(df):
    """
    Validate a Bayesian network for structural and probabilistic consistency.

    Args:
        df: DataFrame containing the extracted network data

    Returns:
        dict: Validation results including errors and warnings
    """
    results = {
        'errors': [],
        'warnings': [],
        'is_valid': True
    }
    # Check for acyclicity
    G = nx.DiGraph()
    for idx, row in df.iterrows():
        title = row['Title']
        parents = row['Parents'] if isinstance(row['Parents'], list) else []

        for parent in parents:
            G.add_edge(parent, title)

    if not nx.is_directed_acyclic_graph(G):
        results['errors'].append("Graph contains cycles and is not a valid DAG")
        results['is_valid'] = False

    # Check for undefined nodes
    all_nodes = set(df['Title'])
    all_parents = set()
    for parents in df['Parents']:
        if isinstance(parents, list):
            all_parents.update(parents)

    undefined_parents = all_parents - all_nodes
    if undefined_parents:
        results['errors'].append(f"Graph contains undefined parent nodes: {undefined_parents
        results['is_valid'] = False

    # Check probability distributions
    for idx, row in df.iterrows():
        title = row['Title']
        priors = row['priors'] if isinstance(row['priors'], dict) else {}

        # Check if prior probabilities sum to 1
        if priors:
            prior_values = []
            for key, value in priors.items():
                if key != 'true_prob':  # Skip derived values
                    try:
                        prior_values.append(float(value))
                    except (ValueError, TypeError):
                        results['warnings'].append(f"Node {title} has non-numeric prior: {va

            if prior_values and abs(sum(prior_values) - 1.0) > 0.01:
                results['warnings'].append(
                    f"Prior probabilities for node {title} sum to {sum(prior_values)}, not 1
                )

        # Additional checks for conditional probabilities omitted for brevity

    return results
```

This validation function performs critical checks for structural integrity (acyclicity, completeness) and probabilistic consistency (distributions summing to 1), identifying errors that would invalidate the network and warnings about potential issues requiring attention.

Error Analysis provides insights into challenging cases and opportunities for improvement. Figure 11 shows a confusion matrix for node relationship classification:

[FIGURE 11: Confusion matrix for node relationship classification]

The confusion matrix reveals that most errors involve failing to detect relationships (false negatives) rather than incorrectly identifying non-existent relationships (false positives). This pattern suggests that the extraction process is conservative, prioritizing precision over recall—generally appropriate for formal modeling where incorrect relationships could lead to substantive errors in reasoning.

The performance and validation assessment demonstrates that the AMTAIR prototype achieves sufficient accuracy and efficiency for practical applications while highlighting specific areas for improvement. The system performs well on structural extraction, shows acceptable but lower accuracy on probability extraction, and handles computational demands efficiently enough for interactive use.

These results validate the fundamental approach while identifying clear paths for enhancement. The two-stage extraction process proves effective for separating structural and probabilistic aspects, with higher performance on the former suggesting that future work should focus particularly on improving probability extraction methods, perhaps through specialized prompting techniques or additional validation mechanisms.

5. Analysis and Results

5.1 Structural Insights from Carlsmith's Model

The formalization of Carlsmith's model reveals structural patterns that might not be apparent from the original text, providing insights into the causal architecture of his argument. By analyzing the network structure mathematically, we can identify key variables, critical pathways, and important dependencies that shape his assessment of existential risk.

One powerful analytical approach examines **centrality measures** that identify influential nodes in the network. Rather than relying on intuition or frequency of mention, these metrics quantify how variables connect to and influence others in the causal structure. Table 2 presents centrality measures for key variables in Carlsmith's model:

Variable	Degree Centrality	Betweenness Centrality	Closeness Centrality
Misaligned_Power_Seeking	0.85	0.42	0.76
Human_Disempowerment	0.35	0.18	0.58
APS_Systems	0.30	0.09	0.45
Scale_Of_Power_Seeking	0.45	0.15	0.64
Existential_Catastrophe	0.15	0.00	0.38

"Misaligned_Power_Seeking" emerges as the most central variable across all metrics, serving as a hub that connects multiple causal pathways. This aligns with Carlsmith's explicit focus on power-seeking behavior as the critical mechanism for existential risk, but the quantitative analysis reveals just how dominant this variable is in the overall structure.

The high betweenness centrality of "Misaligned_Power_Seeking" (0.42) indicates that it serves as a bridge between different clusters of variables. Changes to this variable would affect multiple pathways simultaneously, making it a critical leverage point for risk reduction. This suggests that interventions targeting misaligned power-seeking behavior specifically (rather than just general AI capabilities or deployment decisions) might have outsized effects on existential risk.

Beyond individual variables, **path analysis** identifies critical causal chains leading to existential catastrophe. The formalized model reveals three distinct pathways:

1. **Technical pathway**: Advanced_AI_Capability → Agentic_Planning → Strategic_Awareness → APS_Systems → Misaligned_Power_Seeking → Scale_Of_Power_Seeking → Human_Disempowerment → Existential_Catastrophe
2. **Governance pathway**: Incentives_To_Build_APS → Deployment_Decisions → Misaligned_Power_Seeking → Scale_Of_Power_Seeking → Human_Disempowerment → Existential_Catastrophe
3. **Correction pathway**: Warning_Shots → Corrective_Feedback → Scale_Of_Power_Seeking → Human_Disempowerment → Existential_Catastrophe

These pathways represent different causal mechanisms through which existential catastrophe might occur, suggesting distinct intervention approaches. The technical pathway emphasizes alignment challenges, the governance pathway focuses on deployment incentives, and the correction pathway highlights societal response capabilities.

Another structural insight comes from **Markov blanket analysis**, which identifies the minimal set of variables needed to shield a node from the rest of the network. For "Existential_Catastrophe," the Markov blanket consists solely of "Human_Disempowerment," indicating that in Carlsmith's model, humanind disempowerment completely mediates all pathways to catastrophe.

Similarly, the Markov blanket for "Misaligned_Power_Seeking" includes:

- Parents: APS_Systems, Difficulty_Of_Alignment, Deployment_Decisions
- Children: Scale_Of_Power_Seeking, Barriers_To_Understanding, Adversarial_Dynamics, Stakes_Of_Error
- Children's other parents: Corrective_Feedback

This set represents the minimal contextual information needed to reason about misaligned power-seeking, highlighting the interdependence between technical factors (APS systems, alignment difficulty), governance decisions, and feedback mechanisms.

The formalization also reveals structural asymmetries in Carlsmith's argument. The variables most proximate to existential catastrophe (Human_Disempowerment, Scale_Of_Power_Seeking) have relatively simple causal structures, while technical factors near the bottom of the causal chain (APS_Systems, Difficulty_Of_Alignment) have more complex structures with multiple parent and child relationships. This suggests that Carlsmith's analysis is more nuanced about technical mechanisms than about how power-seeking ultimately leads to catastrophe.

Visual network analysis provides additional insights. Figure 12 shows a force-directed layout of Carlsmith's model with nodes sized according to their betweenness centrality:

[FIGURE 12: Force-directed layout of Carlsmith's model with nodes sized by centrality]

This visualization reveals three distinct clusters in the network:

1. A technical cluster focused on AI capabilities and alignment challenges
2. A governance cluster centered on deployment decisions and incentives
3. A consequences cluster linking power-seeking to ultimate outcomes

The formalized model also enables more sophisticated structural analyses using established network algorithms:

```
def analyze_network_structure(G):
    """
    Perform structural analysis on a Bayesian network.

    Args:
        G: NetworkX DiGraph representing the Bayesian network

    Returns:
        dict: Analysis results including centrality measures,
              communities, and critical paths
    """
    results = {}

    # Calculate centrality measures
    results['degree_centrality'] = nx.degree_centrality(G)
    results['betweenness_centrality'] = nx.betweenness_centrality(G)
    results['closeness_centrality'] = nx.closeness_centrality(G)

    # Identify communities
    undirected_G = G.to_undirected()
    communities = list(nx.community.greedy_modularity_communities(undirected_G))
    results['communities'] = communities

    # Find critical paths
    target_node = 'Existential_Catastrophe'
    if target_node in G.nodes():
        # Find all simple paths to target
        all_paths = []
        for node in G.nodes():
            if node != target_node:
                paths = list(nx.all_simple_paths(G, node, target_node))
                all_paths.extend(paths)

        # Sort paths by length
        all_paths.sort(key=len)
        results['critical_paths'] = all_paths

    return results
```

This function implements various network analysis techniques to extract structural insights, including community detection that identifies clusters of tightly connected variables and critical path analysis that finds all causal chains leading to existential catastrophe.

Perhaps the most valuable structural insight is the identification of “Misaligned_Power_Seeking” as the central hub of Carlsmith’s model. This variable not only has the highest centrality measures but also connects multiple causal pathways, suggesting that it represents a critical junction where technical, governance, and societal factors converge. This aligns with Carlsmith’s explicit focus but quantifies its central role in his analysis.

The structural analysis also reveals potential blindspots or simplifications in Carlsmith’s model. For example, the relatively simple path from “Human_Disempowerment” to “Existential_Catastrophe" suggests limited exploration of how exactly disempowerment leads to catastrophic outcomes. Similarly, the limited connections between technical and governance clusters might indicate insufficient attention to how these domains interact in practice.

These structural insights demonstrate the value of formalization beyond mere representation. By making implicit patterns explicit, the formalized model enables identification of central variables, critical pathways, and structural properties that might not be apparent from the original text. These insights can guide further research, highlight areas for model refinement, and inform intervention strategies focused on the most influential components of the causal structure.

5.2 Probabilistic Assessment and Sensitivity

Beyond structural insights, formalizing Carlsmith's model enables probabilistic analysis that examines the quantitative implications of his judgments. The Bayesian network representation allows calculation of joint and conditional probabilities, sensitivity analysis of critical parameters, and uncertainty propagation through the causal structure.

The first question many readers might ask is: does the formalized model replicate Carlsmith's bottom-line assessment? His paper concludes with approximately 5% probability of existential catastrophe from power-seeking AI, derived from multiplying probabilities across his six key premises. The Bayesian network calculation yields 4.98%, remarkably close to his stated estimate despite the formalization capturing many more details and dependencies.

This agreement validates the formalization approach, demonstrating that the Bayesian network accurately represents Carlsmith's probabilistic judgments. However, the formalized model goes beyond replication to enable more sophisticated analyses.

Sensitivity analysis identifies which parameters most significantly affect the probability of existential catastrophe. By systematically varying individual probabilities and observing the change in the outcome, we can determine which factors have the greatest influence on the bottom-line assessment. Table 3 shows sensitivity results for key variables:

Variable	Baseline State	Alternative State	Change in P(Doom)	Sensitivity Coefficient
Misaligned_Power_Seeking	P(TRUE) = 0.338	P(TRUE) = 0.438	+2.92%	0.292
Corrective_Feedback	P(EFFECTIVE) = 0.60	P(EFFECTIVE) = 0.70	-1.86%	0.186
Deployment_Decisions	P(DEPLOY) = 0.70	P(DEPLOY) = 0.60	-1.67%	0.167
Difficulty_Of_Alignment	P(TRUE) = 0.40	P(TRUE) = 0.50	+1.43%	0.143

Variable	Baseline State	Alternative State	Change in P(Doom)	Sensitivity Coefficient
Advanced_AI_Capability	P(TRUE) = 0.80	P(TRUE) = 0.90	+0.61%	0.061

The sensitivity coefficient represents the rate of change in the probability of existential catastrophe relative to the change in the variable's probability. Higher coefficients indicate greater influence on the outcome.

"Misaligned_Power_Seeking" emerges as the most sensitive variable, with a 10 percentage point increase in its probability causing a 2.92 percentage point increase in existential catastrophe probability. This aligns with its central structural position but quantifies its influence in probabilistic terms.

Interestingly, "Corrective_Feedback" shows the second-highest sensitivity, with increased effectiveness substantially reducing catastrophe probability. This suggests that society's ability to detect and respond to warning signs might be more important than previously recognized, potentially shifting intervention priorities.

The sensitivity analysis can be implemented with the following code:

```python
def sensitivity_analysis(model, target_node, target_state, parameters):
    """
    Perform sensitivity analysis on a Bayesian network.

    Args:
        model: Bayesian network model
        target_node: Outcome variable to measure
        target_state: State of the outcome variable
        parameters: List of (node, state, baseline, alternative) tuples

    Returns:
        dict: Sensitivity results for each parameter
    """
    inference = VariableElimination(model)

    # Get baseline probability
    baseline_query = inference.query(variables=[target_node])
    baseline_prob = baseline_query.values[baseline_query.state_names[target_node].index(targe

    results = {}

    # Test each parameter
    for node, state, baseline_value, alternative_value in parameters:
        # Store original CPD
        original_cpd = model.get_cpds(node)

        # Create modified CPD
        # Implementation details omitted for brevity

        # Replace CPD in model
        model.remove_cpds(original_cpd)
        model.add_cpds(modified_cpd)

        # Calculate new probability
        modified_query = inference.query(variables=[target_node])
        modified_prob = modified_query.values[modified_query.state_names[target_node].index(

        # Calculate sensitivity
        absolute_change = modified_prob - baseline_prob
        relative_change = absolute_change / (alternative_value - baseline_value)

        results[node] = {
            'baseline_prob': baseline_prob,
            'modified_prob': modified_prob,
            'absolute_change': absolute_change,
            'sensitivity_coefficient': relative_change
        }

        # Restore original CPD
        model.remove_cpds(modified_cpd)
        model.add_cpds(original_cpd)

    return results
```

This function implements sensitivity analysis by systematically modifying individual parameters, calculating the resulting change in outcome probability, and computing sensitivity coefficients. The approach maintains model integrity by restoring original parameters after each test.

Beyond individual parameter sensitivity, the formalized model enables **uncertainty propagation** analysis that examines how parameter uncertainty affects conclusions. Instead of using point estimates for probabilities, we can represent each parameter as a probability distribution reflecting our uncertainty about its true value. These distributions propagate through the network, creating a distribution over the probability of existential catastrophe.

Figure 13 shows the result of uncertainty propagation for Carlsmith's model:

[FIGURE 13: Probability distribution over P(Doom) reflecting parameter uncertainty]

The distribution has a mean of 4.98% (matching Carlsmith's estimate) but spans from approximately 1% to 12%, reflecting uncertainty in the underlying parameters. This analysis suggests that while Carlsmith's central estimate is reasonable, the true probability could be substantially higher or lower depending on parameter values.

The uncertainty propagation highlights an important aspect of existential risk assessment: precise probability estimates can create false precision, while ranges better represent our actual state of knowledge. The formalized model makes this uncertainty explicit, enabling more nuanced discussion of risk levels.

Another valuable probabilistic analysis examines **conditional relationships** between variables, revealing how different factors interact to influence outcomes. For example, we can calculate the probability of existential catastrophe under different combinations of "Corrective_Feedback" and "Deployment_Decisions":

Corrective_Feedback	Deployment_Decisions	P(Existential_Catastrophe)
EFFECTIVE	DEPLOY	3.74%
EFFECTIVE	WITHHOLD	1.52%
INEFFECTIVE	DEPLOY	7.33%
INEFFECTIVE	WITHHOLD	3.87%

This analysis reveals interesting interactions: effective corrective feedback reduces catastrophe probability by approximately 50% regardless of deployment decisions, while withholding deployment reduces probability by approximately 60% regardless of feedback effectiveness. The combination of both interventions (effective feedback and withholding deployment) reduces probability by nearly 80% compared to the worst case.

Such conditional analyses enable more sophisticated reasoning about intervention combinations, identifying synergies between different approaches rather than focusing on individual factors in isolation.

The probabilistic assessments provide several key insights:

1. Carlsmith's bottom-line estimate of approximately 5% probability for existential catastrophe is correctly replicated in the formalized model, validating the formalization approach.

2. Misaligned power-seeking emerges as both structurally central and highly sensitive, confirming its critical role in the risk pathway.

3. Corrective feedback appears more important than initially apparent, suggesting increased attention to societal response mechanisms.

4. Parameter uncertainty creates substantial variation in the bottom-line estimate, highlighting the importance of ranges rather than point estimates.
5. Interventions display interesting interaction effects, with combinations potentially offering greater risk reduction than the sum of individual approaches.

These insights demonstrate the value of formalization for probabilistic reasoning. By making relationships and judgments explicit in a computational framework, the formalized model enables sophisticated analyses that reveal patterns, sensitivities, and implications not obvious from the original text.

5.3 Policy Impact Evaluation

Moving beyond structural and probabilistic analysis, the formalized Carlsmith model enables systematic evaluation of how governance interventions might affect existential risk. This capability bridges theoretical understanding and practical action, allowing policymakers to explore the potential consequences of different approaches before implementation.

Policy impact evaluation in the AMTAIR system uses counterfactual analysis based on Pearl's do-calculus, which distinguishes between observing and intervening on variables. Rather than simply calculating conditional probabilities (what happens if we observe a variable taking a certain value), the system models interventions that force variables to specific values and propagates these changes through the causal structure.

To demonstrate this approach, I modeled several candidate policies as interventions on specific variables in Carlsmith's model:

1. **Mandatory safety demonstrations** require developers to prove alignment properties before deployment, modeled as an intervention on "Deployment_Decisions" to increase the probability of withholding misaligned systems.
2. **Compute governance frameworks** restrict access to computational resources needed for advanced AI training, modeled as an intervention reducing the probability of "Advanced_AI_Capability" reaching transformative levels.
3. **Monitoring and feedback systems** enhance detection of and response to warning signs from early systems, modeled as an intervention increasing the probability of "Warning_Shots" being observed and "Corrective_Feedback" being effective.

The system implements these interventions through manipulations of the Bayesian network structure and parameters:

```
def implement_policy_intervention(model, intervention_type, params):
    """
    Implement a policy intervention on the Bayesian network.

    Args:
        model: The Bayesian network model
        intervention_type: Type of intervention ('do', 'soft', 'mechanism')
        params: Parameters specific to the intervention type

    Returns:
        Modified model with intervention implemented
    """
    # Create a copy of the model to avoid modifying the original
    modified_model = model.copy()

    if intervention_type == 'do':
        # Hard intervention setting variable to specific value
        variable = params['variable']
        value = params['value']

        # Remove all incoming edges to the variable
        parents = list(modified_model.get_parents(variable))
        for parent in parents:
            modified_model.remove_edge(parent, variable)

        # Set new CPD with certainty for the specified value
        old_cpd = modified_model.get_cpds(variable)
        variable_card = old_cpd.variable_card
        state_names = old_cpd.state_names[variable]

        # Create values array with 1.0 for specified value, 0.0 for others
        values = np.zeros((variable_card, 1))
        target_idx = state_names.index(value)
        values[target_idx] = 1.0

        # Create new CPD and add to model
        new_cpd = TabularCPD(
            variable=variable,
            variable_card=variable_card,
            values=values,
            state_names={variable: state_names}
        )
        modified_model.remove_cpds(old_cpd)
        modified_model.add_cpds(new_cpd)

    elif intervention_type == 'soft':
        # Soft intervention modifying probability distribution
        variable = params['variable']
        distribution = params['distribution']

        # Keep existing structure but modify CPD
        old_cpd = modified_model.get_cpds(variable)
        parents = list(modified_model.get_parents(variable))

        if not parents:
            # For root nodes, simply replace distribution
            variable_card = old_cpd.variable_card
            state_names = old_cpd.state_names[variable]

            # Create values array from new distribution
            values = np.array([distribution]).T

            # Create new CPD and add to model
            new_cpd = TabularCPD(
                variable=variable,
                variable_card=variable_card,
                values=values,
                state_names={variable: state_names}
            )

            modified_model.remove_cpds(old_cpd)
            modified_model.add_cpds(new_cpd)
        else:
            # For nodes with parents, modify CPD while preserving structure
            # Implementation details omitted for brevity

    elif intervention_type == 'mechanism':
        # Mechanism intervention modifying causal structure
        add_edges = params.get('add_edges', [])
        remove_edges = params.get('remove_edges', [])
        modify_cpds = params.get('modify_cpds', {})

        # Add and remove edges as specified
        for source, target in remove_edges:
            if modified_model.has_edge(source, target):
                modified_model.remove_edge(source, target)

        for source, target in add_edges:
            if not modified_model.has_edge(source, target):
                modified_model.add_edge(source, target)

        # Modify CPDs as specified
        for variable, cpd_params in modify_cpds.items():
            # Implementation details omitted for brevity

    # Verify the modified model is valid
    modified_model.check_model()

    return modified_model
```

This function supports three types of interventions:

1. **Hard interventions** (do-operations) force variables to specific values by removing incoming edges and setting fixed probabilities
2. **Soft interventions** modify probability distributions while preserving the existing causal structure
3. **Mechanism interventions** change the causal structure itself by adding or removing edges between variables

These different intervention types enable modeling various policy approaches, from direct regulations that force specific behaviors to incentive structures that modify probabilities without guaranteeing outcomes.

To evaluate the safety demonstrations policy, I implemented a soft intervention on "Deployment_Decisions" that reduced the probability of deploying potentially misaligned systems:

```
# Define the safety demonstrations policy
safety_demo_policy = {
    'intervention_type': 'soft',
    'params': {
        'variable': 'Deployment_Decisions',
        'distribution': {
            'p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_T
            'p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_FA
            'p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_TRU
            'p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_FALS
            'p(deployment_decisions_WITHHOLD|incentives_to_build_aps_STRONG, deception_by_ai_
            'p(deployment_decisions_WITHHOLD|incentives_to_build_aps_STRONG, deception_by_ai_
            'p(deployment_decisions_WITHHOLD|incentives_to_build_aps_WEAK, deception_by_ai_T
            'p(deployment_decisions_WITHHOLD|incentives_to_build_aps_WEAK, deception_by_ai_F
        }
    }
}

# Implement the policy and evaluate impact
modified_model = implement_policy_intervention(carlsmith_network, **safety_demo_policy)
safety_impact = evaluate_policy_impact(
    modified_model,
    target_variable='Existential_Catastrophe',
    target_value='existential_catastrophe_TRUE'
)
```

This code implements the safety demonstrations policy by modifying the conditional probabilities for “Deployment_Decisions” based on its parent variables (“Incentives_To_Build_APS” and “Deception_By_AI”), reducing the probability of deployment across all parent combinations.

The evaluation results showed that safety demonstrations reduced the probability of existential catastrophe from 4.98% to 3.21%, a 35.5% relative reduction. This significant impact highlights the potential value of governance measures targeting deployment decisions.

I applied similar analyses to the compute governance and monitoring policies, with results summarized in Table 4:

Policy	Implementation	Baseline P(Doom)	Intervention P(Doom)	Absolute Change	Relative Change
Safety Demonstrations	Modify Deployment_Decisions	4.98%	3.21%	-1.77%	-35.5%
Compute Governance	Modify Advanced_AI_Capability	4.98%	4.10%	-0.88%	-17.7%
Monitoring & Feedback	Modify Warning_Shots & Corrective_Feedback	4.98%	2.74%	-2.24%	-45.0%

Interestingly, the monitoring and feedback policy showed the largest impact despite modifying variables further from “Existential_Catastrophe” in the causal chain. This suggests that enhancing society's ability to detect and respond to early warning signs might be more effective than directly regulating deployment or restricting compute access.

To explore this further, I compared the policies across different parameter variations to assess their robustness. Figure 14 shows how each policy performs under variations in the probability of "Difficulty_Of_Alignment":

[FIGURE 14: Graph showing policy effectiveness across different alignment difficulty scenarios]

The monitoring and feedback policy maintained the largest impact across all scenarios, while compute governance showed diminishing effectiveness as alignment difficulty increased. This suggests that monitoring systems might provide more robust risk reduction regardless of how difficult alignment proves to be.

Beyond evaluating individual policies, the system enables assessment of policy portfolios—combinations of interventions that might create synergistic effects. I modeled a comprehensive governance framework combining all three policies:

```python
# Define comprehensive governance framework
comprehensive_framework = {
    'safety_demonstrations': safety_demo_policy,
    'compute_governance': compute_gov_policy,
    'monitoring_feedback': monitoring_policy
}

# Implement all policies sequentially
comprehensive_model = carlsmith_network.copy()
for policy_name, policy_params in comprehensive_framework.items():
    comprehensive_model = implement_policy_intervention(
        comprehensive_model,
        policy_params['intervention_type'],
        policy_params['params']
    )

# Evaluate combined impact
comprehensive_impact = evaluate_policy_impact(
    comprehensive_model,
    target_variable='Existential_Catastrophe',
    target_value='existential_catastrophe_TRUE'
)
```

The comprehensive framework reduced the probability of existential catastrophe from 4.98% to 1.32%, a 73.5% relative reduction. Notably, this exceeds the sum of individual policy impacts

(4.89% combined absolute reduction), suggesting synergistic effects where policies complement each other.

These results demonstrate the value of formal policy evaluation for existential risk governance. By modeling interventions in a structured causal framework, we can:

1. Quantify expected impacts on risk levels based on explicit assumptions
2. Compare different governance approaches on a common basis
3. Identify unexpectedly high-leverage intervention points
4. Assess policy robustness across different parameter variations
5. Evaluate complementarities between different policies

This capability addresses a critical gap in current governance discussions, where debates often focus on abstract principles rather than expected outcomes. The formalized approach enables more concrete conversations about causal mechanisms, effect magnitudes, and intervention designs.

It's important to note that these evaluations depend on the causal structure and probability values encoded in Carlsmith's model. Different models might yield different policy evaluations, highlighting the importance of making assumptions explicit and testing interventions across multiple worldviews. The AMTAIR approach facilitates this kind of cross-worldview assessment by applying the same evaluation methodology to different formalized models.

5.4 Cross-Domain Integration Potential

Beyond the specific analytical capabilities demonstrated in previous sections, the AMTAIR approach offers broader potential for integrating insights and coordinating efforts across the disparate domains involved in AI governance. This section explores how the approach bridges technical and policy communities, enhances cross-stakeholder understanding, and supports strategic coordination.

The coordination gap in AI governance stems partly from communication barriers between domains. Technical AI alignment researchers often express concerns in mathematical formalisms inaccessible to policy specialists. Governance experts frequently frame issues in institutional terms unfamiliar to technical researchers. Ethicists articulate principles without operational details for implementation. Each domain develops sophisticated insights, but these remain siloed without effective integration mechanisms.

The AMTAIR approach addresses this gap by creating shared representations that maintain connections to domain-specific knowledge while enabling cross-domain communication. The system creates bridges along several dimensions:

1. Technical-policy integration connects technical alignment research with governance frameworks by representing both in a common causal structure. Technical factors like instrumental convergence and proxy optimization become nodes in the same network as governance factors like deployment decisions and institutional oversight, making their relationships explicit. This connection enables bidirectional influence:

- Technical insights inform governance by showing how alignment challenges affect risk pathways, helping prioritize regulations based on their causal impact.
- Governance perspectives inform technical work by highlighting institutional constraints and implementation pathways, guiding research toward solutions with practical application.

The formalization of Carlsmith's model demonstrates this integration, representing both technical factors (e.g., Advanced_AI_Capability, Problems_With_Proxies) and governance considerations (e.g., Deployment_Decisions, Corrective_Feedback) in a unified causal structure. The analysis reveals how these domains interact—for example, showing how effective corrective feedback can partially mitigate misaligned power-seeking, creating resilience against technical failures.

2. Research prioritization insights emerge from analyzing formalized models to identify high-leverage variables and critical uncertainties. By examining sensitivity and centrality, the system identifies which factors most significantly influence risk levels, helping direct research efforts toward areas with the greatest potential impact.

The analysis of Carlsmith's model revealed that "Misaligned_Power_Seeking" combines high centrality and sensitivity, suggesting research prioritization for:

- Technical approaches that reduce the likelihood of misalignment leading to power-seeking behavior
- Governance mechanisms that detect and respond to early signs of misaligned power-seeking
- Monitoring systems that track indicators of power-seeking behavior in AI systems

These priorities span technical and governance domains, guiding collaborative research efforts that integrate multiple perspectives rather than pursuing siloed approaches.

3. Communication enhancement through intuitive visualizations and progressive disclosure makes complex models accessible to diverse stakeholders. The interactive visualization system presents information at multiple levels of detail, allowing individuals to engage based on their background and interests:

- Technical experts can explore detailed probability distributions and sensitivity analyses
- Policy specialists can focus on intervention impacts and governance pathways
- Generalists can understand overall structure and key relationships

This multi-level accessibility helps bridge the "formalism barrier" that often prevents non-technical stakeholders from engaging with formal models. Rather than requiring all participants to adopt a single specialized language, the system provides multiple entry points while maintaining a consistent underlying representation.

4. Implementation pathways become clearer by connecting abstract governance principles to concrete causal mechanisms. The policy evaluation capability demonstrates how specific interventions influence risk through particular causal pathways, helping translate high-level goals into operational details:

```python
def map_governance_principles_to_mechanisms(principles, model):
    """
    Map high-level governance principles to specific causal mechanisms.

    Args:
        principles: List of governance principles
        model: Bayesian network representing causal structure

    Returns:
        dict: Mapping from principles to causal mechanisms
    """
    mapping = {}

    for principle in principles:
        # Identify variables influenced by this principle
        affected_variables = identify_affected_variables(principle, model)

        # Determine potential intervention types
        intervention_options = []
        for variable in affected_variables:
            # Analyze causal structure to determine appropriate interventions
            variable_parents = list(model.get_parents(variable))
            variable_children = list(model.get_children(variable))

            # Suggest intervention types based on variable's position in causal structure
            if not variable_parents:  # Root cause
                intervention_options.append({
                    'variable': variable,
                    'type': 'direct_modification',
                    'mechanism': f"Directly modify {variable} distribution"
                })
            else:  # Intermediate variable
                intervention_options.append({
                    'variable': variable,
                    'type': 'structural_change',
                    'mechanism': f"Modify relationship between {variable_parents} and {varial
                })
                intervention_options.append({
                    'variable': variable,
                    'type': 'conditional_modification',
                    'mechanism': f"Modify {variable} distribution conditional on {variable_pa
                })

        mapping[principle] = {
            'affected_variables': affected_variables,
            'intervention_options': intervention_options
        }

    return mapping
```

This function helps bridge abstract principles and concrete mechanisms by identifying which variables in the causal model relate to specific governance principles and suggesting appropriate intervention types based on the variables' positions in the causal structure. This mapping helps translate high-level goals into specific implementation details.

5. Integration with existing frameworks helps connect the formalized approach to current governance initiatives. Rather than creating a parallel system, the AMTAIR approach complements existing frameworks by enhancing their analytical foundations:

- **Technical standards development** benefits from explicit causal models showing how different technical properties influence risk pathways, informing standard scope and validation criteria.
- **Regulatory frameworks** gain precision through formal analysis of how specific regulations affect causal mechanisms, enhancing impact assessment and identifying potential unintended consequences.
- **Multi-stakeholder initiatives** benefit from shared representations that make implicit assumptions explicit, facilitating more productive discourse across different perspectives.

The potential for integration extends to specific organizations and initiatives such as the Partnership on AI, NIST AI Risk Management Framework, and OECD AI Principles. Each of these efforts could enhance its analytical foundation through formalized models that make causal assumptions explicit and enable systematic comparison across perspectives.

6. Adoption considerations influence how the approach translates from research prototype to practical implementation. Several factors affect potential adoption:

- **Accessibility barriers** due to technical complexity could limit participation without appropriate interfaces and documentation.

- **Institutional incentives** might resist formalization that makes implicit assumptions explicit and subject to critique.
- **Resource requirements** for model development and maintenance might constrain adoption without adequate funding and organizational support.
- **Integration with existing processes** requires thoughtful design to avoid creating parallel systems that increase rather than reduce coordination burden.

The implementation approach would need to address these considerations through incremental deployment, stakeholder co-design, and integration with existing workflows rather than wholesale replacement of current processes.

Despite these challenges, the cross-domain integration potential of the AMTAIR approach addresses a fundamental need in AI governance: coordinating diverse efforts toward coherent strategies for managing existential risk. By creating shared representations that bridge technical and policy domains, making implicit models explicit, and enabling systematic comparison across perspectives, the approach provides crucial infrastructure for the kind of coordination necessary as AI capabilities continue to advance.

The ultimate vision is not a single, authoritative model that all stakeholders must adopt, but rather an ecosystem of interoperable models that retain domain-specific knowledge while enabling cross-domain communication and integration. This ecosystem would support more effective coordination by making assumptions explicit, facilitating structured comparison, and identifying genuine points of agreement and disagreement across perspectives.

6. Counterclaims and Rebuttals

6.1 Formalization Limitations

COUNTERCLAIM: Formal models inherently oversimplify complex governance challenges, stripping away critical context and nuance. By reducing rich qualitative arguments to nodes, edges, and probability distributions, the AMTAIR approach loses the depth and context of original reasoning. This oversimplification can create false precision and misguided confidence, potentially leading to worse governance decisions than qualitative approaches grounded in contextual understanding.

This perspective has merit in several contexts. The history of policy analysis contains numerous examples where formalization led to detrimental outcomes. During the Vietnam War, Secretary of Defense Robert McNamara's systems analysis approach applied quantitative optimization to warfare, using metrics like "body counts" that distorted military strategy and ignored crucial cultural and political factors. Similarly, economic models that reduced complex financial systems to simplified mathematical relationships contributed to the 2008 financial crisis by creating overconfidence in risk management capabilities.

In AI governance specifically, formal models might oversimplify value alignment challenges by reducing complex normative considerations to simple utility functions. They might miss important sociocultural factors that influence how technologies are developed and deployed across different contexts. And they could create false certainty about causal relationships that remain deeply uncertain and contingent.

REBUTTAL: Appropriate formalization enhances rather than replaces qualitative understanding by making implicit assumptions explicit and enabling structured reasoning about complex relationships. The AMTAIR approach specifically addresses oversimplification concerns through hybrid representations that preserve narrative context alongside formal structure.

First, BayesDown maintains natural language descriptions alongside formal elements, preserving the qualitative reasoning that informs the model. Unlike purely mathematical formalisms that strip away context, BayesDown retains descriptions for each variable and relationship, maintaining connections to the original arguments and allowing users to refer back to qualitative reasoning.

Second, the interactive visualization system provides progressive disclosure of information, allowing users to explore both structural patterns and narrative details. The layered approach presents simple causal diagrams for initial understanding, with deeper exploration revealing detailed probability information and qualitative descriptions. This maintains complexity while making it navigable.

Third, uncertainty representation is explicit throughout the system, with probability distributions rather than point estimates and sensitivity analysis that reveals which factors significantly influence outcomes. Far from creating false precision, this approach makes uncertainty visible and quantifiable, enhancing epistemic humility rather than diminishing it.

Consider how these features manifest in the Carlsmith model implementation:

```
[Existential_Catastrophe]: The destruction of humanity's long-term potential due to AI system
  "instantiations": ["existential_catastrophe_TRUE", "existential_catastrophe_FALSE"],
  "priors": {"p(existential_catastrophe_TRUE)": "0.05", "p(existential_catastrophe_FALSE)": "
  "posteriors": {
    "p(existential_catastrophe_TRUE|human_disempowerment_TRUE)": "0.95",
    "p(existential_catastrophe_TRUE|human_disempowerment_FALSE)": "0.0"
  }
}
```

This representation maintains Carlsmith's original description of existential catastrophe ("The destruction of humanity's long-term potential…") alongside the formal structure and probability information. The qualitative reasoning remains accessible, providing context for the quantitative elements.

The interactive visualization similarly preserves qualitative content, with tooltips showing descriptions and expanded views providing detailed narrative context. This maintains connections to original reasoning while enabling formal analysis.

SYNTHESIS: A balanced approach recognizes formalization's value while acknowledging its limitations. Rather than choosing between formal models and qualitative reasoning, effective governance analysis integrates both, using models to structure thinking while maintaining narrative context and domain expertise.

The appropriate approach involves:

1. **Complementary use** of formal and qualitative methods, with models supporting rather than replacing expert judgment
2. **Transparent assumptions** that make formalization choices explicit and subject to critique
3. **Iterative refinement** based on stakeholder feedback and evolving understanding
4. **Domain-appropriate abstraction** that formalizes aspects where formal reasoning adds value while preserving qualitative analysis for others
5. **Contextual presentation** that connects formal results to their qualitative implications

This balanced approach characterizes the AMTAIR system, which uses formalization to enhance cross-domain coordination while maintaining connections to qualitative reasoning and domain expertise. Rather than creating a conflict between formal and qualitative approaches, it establishes bridges between them, enabling analysts to move between levels of abstraction as appropriate for different questions and contexts.

6.2 Epistemic Humility Considerations

COUNTERCLAIM: Quantitative models create false precision and overconfidence in domains characterized by deep uncertainty. By assigning specific probability values to highly uncertain events, the AMTAIR approach might convey unwarranted certainty about AI risk pathways and intervention effects. This numeric precision can create an illusion of knowledge, leading to overconfidence in governance decisions and underestimation of fundamental uncertainties about how advanced AI will develop and impact society.

This concern has substantial historical support. Expert quantitative models have repeatedly led to overconfidence with serious consequences. Long-Term Capital Management, a hedge fund using sophisticated mathematical models developed by Nobel laureates, collapsed in 1998 after its models failed to account for scenario uncertainties outside their historical data. The financial crisis of 2008 stemmed partly from risk models that assigned precise probabilities to mortgage default scenarios without adequately accounting for systemic uncertainties and interdependencies.

In AI governance specifically, we face even deeper uncertainties than these historical examples. We lack empirical data on how transformative AI capabilities might develop, how misalignment might manifest at advanced capability levels, or how institutions might respond to unprecedented challenges. Assigning specific probabilities to these deeply uncertain events might create an unwarranted sense of knowledge about fundamentally unpredictable developments.

REBUTTAL: Explicit representation of uncertainty enhances epistemic humility by making limitations visible rather than implicit. The AMTAIR approach specifically incorporates uncertainty in multiple dimensions—parameter ranges, model structure alternatives, and sensitivity analysis—creating greater awareness of knowledge limitations rather than obscuring them.

First, probability distributions rather than point estimates represent uncertain parameters, acknowledging ranges of plausible values. The uncertainty propagation analysis demonstrated how parameter uncertainty creates a distribution over existential risk probabilities (spanning from approximately 1% to 12% in Carlsmith's model), making uncertainty explicit rather than hidden.

Second, sensitivity analysis quantifies which variables most significantly affect outcomes, highlighting areas of critical uncertainty that warrant particular attention. Rather than concealing the impact of uncertain parameters, this approach makes it explicit and actionable, directing attention to high-leverage uncertainties.

Third, cross-worldview comparison capabilities enable evaluation of interventions across different causal models, acknowledging structural uncertainty about how factors interrelate. This capability supports robust decision-making under model uncertainty, identifying interventions that work reasonably well across different plausible models rather than assuming a single correct representation.

Fourth, the interactive visualization encodes uncertainty through visual elements and progressive disclosure, avoiding presentation styles that imply false precision. The system uses features like graduated color scales, explicit confidence intervals, and narrative descriptions of uncertainty to maintain appropriate epistemic humility.

Research on reasoning under uncertainty supports this approach. Studies show that explicit quantification of uncertainty often reduces overconfidence compared to qualitative judgments, where vague terms like "likely" or "unlikely" mask substantial disagreement about probabilities. Making assumptions explicit, even with approximate probabilities, enables more productive discourse about uncertainties than leaving them implicit in qualitative language.

SYNTHESIS: Balancing quantification with appropriate humility requires thoughtful practices to maintain awareness of fundamental uncertainties while benefiting from structured reasoning. The key is not avoiding quantification but implementing it in ways that enhance rather than diminish epistemic humility.

Effective approaches include:

1. **Representing uncertainty at multiple levels** (parameters, structure, outcomes) rather than focusing solely on parameter uncertainty
2. **Using ranges and distributions** rather than point estimates to avoid false precision
3. **Conducting sensitivity analysis** to identify critical uncertainties that warrant particular attention
4. **Testing interventions across multiple models** to identify robust approaches under structural uncertainty
5. **Combining quantitative and qualitative approaches** to leverage the strengths of both
6. **Maintaining iteration and adaptation** as new information emerges, rather than treating models as fixed representations

The AMTAIR approach implements these practices through its multiple uncertainty representations, sensitivity analysis capabilities, and interactive visualizations that maintain appropriate epistemic humility while enabling structured reasoning about uncertain futures.

This balanced approach recognizes that the choice isn't between quantification and humility, but rather between implicit and explicit uncertainty. By making uncertainties explicit and analyzing their implications systematically, the AMTAIR approach enhances epistemic humility while enabling more rigorous governance analysis.

6.3 Democratic Governance Concerns

COUNTERCLAIM: Technical formalization may exclude stakeholders by creating barriers to participation based on specialized expertise. The AMTAIR approach risks concentrating power among technical experts who can understand and manipulate the formal models, while marginalizing stakeholders without technical backgrounds. This exclusion undermines democratic governance principles requiring broad participation in decisions with significant societal implications, potentially leading to technocratic governance that fails to incorporate diverse perspectives and values.

This concern aligns with broader critiques of expert-driven governance. Technical complexity has often served as a barrier to participation in domains from environmental regulation to financial oversight, where specialized languages and methodologies limit meaningful involvement to those with specific expertise. Even with good intentions, technical approaches can create "black boxes" that resist public scrutiny and accountability.

For AI governance specifically, formalization might exclude important perspectives:

- **Civil society organizations** without technical resources might struggle to engage with formal models

- **Global South stakeholders** with different resources and priorities might have limited influence
- **Diverse public perspectives** that aren't readily formalized might be undervalued
- **Humanistic and ethical considerations** might be reduced to simplified parameters

This exclusion could lead to governance frameworks that reflect narrow technical perspectives while failing to incorporate broader societal values and concerns, ultimately undermining legitimacy and effectiveness.

REBUTTAL: Visualization and interactive exploration enhance rather than reduce accessibility by making complex models interpretable to diverse stakeholders. The AMTAIR approach specifically addresses accessibility through multi-level interfaces, progressive disclosure, and visual encoding that enable engagement without requiring specialized expertise.

First, the interactive visualization system provides multiple entry points based on user background and interests. The basic causal structure uses intuitive visual metaphors (nodes, edges, colors) that require minimal technical understanding, while allowing progressive exploration for those seeking deeper details. This tiered approach enables participation across different expertise levels.

Second, natural language descriptions maintain connections to ordinary language rather than requiring specialized vocabulary. The BayesDown format preserves narrative descriptions alongside formal elements, and the visualization displays these descriptions prominently, maintaining accessibility for non-technical users.

Third, visual encoding of probability through color gradients and interactive elements makes quantitative information intuitively understandable without requiring statistical expertise. Rather than presenting complex mathematical notations or tables of numbers, the visualization uses visual metaphors that align with natural cognitive processes.

Fourth, the system supports collective exploration by making models shareable and accessible through standard web browsers, enabling distributed analysis across different stakeholder groups. This accessibility supports collaborative examination and critique rather than isolated technical analysis.

Research on participatory modeling and visualization supports this approach. Studies show that appropriate visualizations can make complex models accessible to diverse stakeholders, enhancing understanding and participation rather than limiting it. Interactive interfaces that allow exploration without requiring model construction can be particularly effective for engaging non-technical participants.

The AMTAIR visualization demonstrates these principles through features like:

- Color-coded nodes that intuitively represent probability values
- Tooltips that reveal additional information on hover
- Modal windows that provide detailed explanations on click
- Interactive layout that allows reorganization based on user interest

- Progressive disclosure that reveals details based on user engagement

These features create bridges between technical formalism and intuitive understanding, enabling participation without requiring specialized expertise.

SYNTHESIS: Designing for inclusive participation while maintaining analytical rigor requires thoughtful approaches that bridge technical and non-technical perspectives. The goal should be enabling meaningful engagement across diverse expertise levels rather than choosing between technical sophistication and accessibility.

Effective approaches include:

1. **Multi-level interfaces** that provide different entry points based on background and interests
2. **Participatory design processes** that incorporate diverse stakeholders in developing visualization approaches
3. **Complementary formats** that present the same information in different ways for different audiences
4. **Capacity building initiatives** that enhance stakeholders' ability to engage with formal models
5. **Bidirectional translation** that moves between technical and non-technical expressions based on context

These approaches recognize that accessibility isn't just about simplifying complex ideas, but about creating appropriate interfaces for different needs and contexts. The AMTAIR visualization implements these principles through its multi-level design, interactive exploration capabilities, and natural language integration.

This balanced approach acknowledges the legitimate concern about technical barriers while demonstrating how thoughtful design can create bridges rather than walls between technical and non-technical stakeholders. By making complex models visually intuitive and progressively explorable, the system enhances democratic participation rather than undermining it.

6.4 Implementation Feasibility

COUNTERCLAIM: The AMTAIR approach faces substantial practical barriers to real-world implementation in governance contexts. Despite theoretical value, the system may prove infeasible in practice due to resource requirements, institutional barriers, adoption challenges, and scaling limitations. Governance institutions often lack technical capacity for sophisticated modeling, face budget constraints limiting investment in novel approaches, and operate under procedural requirements that resist methodological innovation. These practical challenges may prevent the approach from achieving meaningful impact regardless of its theoretical merits.

This concern reflects realistic assessment of implementation barriers. Government agencies and international organizations typically face resource constraints that limit adoption of novel methods, especially those requiring specialized expertise. Many struggle with basic digital infrastructure, let alone advanced modeling capabilities. Institutional processes often evolve slowly through incremental change rather than adopting fundamentally new approaches.

Previous attempts to introduce formal modeling into governance processes illustrate these challenges. The Office of Technology Assessment provided formal analysis to the U.S. Congress but was ultimately defunded despite recognized value. Various environmental modeling initiatives have struggled to achieve sustained adoption in policy processes despite clear relevance. Current AI governance institutions show similar constraints in technical capacity and methodological flexibility.

Specific implementation barriers include:

- **Expertise requirements** for model development and maintenance
- **Data limitations** constraining model validation and calibration
- **Institutional inertia** favoring established methods
- **Integration challenges** with existing decision processes
- **Scaling difficulties** for complex, real-world models

Given these obstacles, even a theoretically valuable approach might fail to achieve practical impact in governance contexts.

REBUTTAL: Incremental implementation paths with progressive enhancement enable practical adoption despite resource constraints and institutional barriers. The AMTAIR approach specifically supports gradual deployment through modular architecture, tiered capability levels, and integration with existing processes.

First, the modular system architecture allows component-wise implementation rather than requiring all-or-nothing adoption. Organizations can begin with basic extraction and visualization capabilities before implementing more sophisticated analysis features, spreading resource requirements over time and allowing incremental value demonstration.

Second, tiered capability levels accommodate different organizational capacities, from simple static visualizations to fully interactive models with live data integration. This tiered approach enables value at various resource levels, with enhanced capabilities available as capacity increases.

Third, integration with existing processes uses familiar interfaces and workflows where possible, reducing adoption barriers and training requirements. Rather than requiring wholesale process changes, the approach can augment existing analysis methods while gradually demonstrating additional value.

Fourth, scalability considerations inform the technical implementation, with attention to computational efficiency and resource requirements. The current prototype demonstrates reasonable performance even on complex models like Carlsmith's, suggesting feasibility for real-world applications.

Historical examples of successful innovation adoption in governance provide instructive parallels. Geographic Information Systems (GIS) initially faced similar barriers but achieved widespread adoption through incremental implementation, starting with basic mapping capabilities before adding sophisticated analysis features. Similarly, economic modeling began with simple tools before expanding to more complex approaches as institutional capacity developed.

A concrete implementation roadmap might include:

1. **Phase 1: Basic Visualization** - Static visualizations of pre-built models requiring minimal technical expertise
2. **Phase 2: Interactive Exploration** - Browser-based interactive visualization with predefined models
3. **Phase 3: Custom Modeling** - Basic extraction and modeling capabilities for specific use cases
4. **Phase 4: Advanced Analysis** - Sensitivity analysis, policy evaluation, and cross-model comparison
5. **Phase 5: Full Integration** - Live data connections, automated updates, and workflow integration

This phased approach distributes resource requirements over time while demonstrating value at each stage, addressing practical adoption constraints.

SYNTHESIS: Realistic implementation requires acknowledging constraints while pursuing feasible adoption paths that deliver incremental value. Rather than seeking immediate comprehensive adoption, effective implementation balances ambition with practicality.

Key principles for successful implementation include:

1. **Value demonstration at each stage** rather than requiring full deployment for initial benefits
2. **Stakeholder engagement** in design and implementation to ensure relevance and usability
3. **Complementary deployment** alongside existing methods rather than immediate replacement
4. **Resource-appropriate configurations** tailored to different organizational contexts

5. **Sustainability planning** for ongoing maintenance and enhancement

The AMTAIR approach supports these principles through its modular architecture, tiered capabilities, and flexible integration options. Rather than presenting an all-or-nothing proposition, it enables progressive enhancement based on resource availability and institutional readiness.

This balanced implementation strategy recognizes legitimate feasibility concerns while identifying practical paths forward. By starting with simpler implementations that deliver immediate value while establishing foundations for more sophisticated capabilities, the approach can achieve meaningful impact despite real-world constraints.

The path forward involves strategic partnerships with organizations like the Partnership on AI, NIST, or the OECD that already engage in AI governance efforts and could integrate AMTAIR capabilities into existing initiatives. These partnerships would provide practical implementation contexts while leveraging established networks for broader adoption.

7. Conclusion and Outlook

7.1 Summary of Key Contributions

This thesis has developed and demonstrated AMTAIR (Automating Transformative AI Risk Modeling), a computational approach that addresses the coordination crisis in AI governance by automating the extraction of probabilistic world models from AI safety literature. The research has made several interrelated contributions that span methodological innovation, technical implementation, analytical capabilities, and governance implications.

Methodologically, the thesis introduced a novel two-stage extraction process that separates structure from probability, improving extraction quality and creating interpretable intermediate representations. The ArgDown format provides a standardized syntax for representing causal structures, while BayesDown extends this to include probabilistic information in a hybrid format that bridges qualitative argumentation and quantitative modeling. This approach aligns with human cognitive processes, first identifying what factors matter and how they relate, then assessing how probable different scenarios are based on those relationships.

Technically, the thesis implemented a complete extraction and analysis pipeline that transforms structured text into interactive Bayesian networks. The implementation processes ArgDown and BayesDown representations into formal network structures, calculates derived properties like centrality measures and Markov blankets, and creates interactive visualizations with probability encoding and progressive disclosure. This pipeline was validated on the canonical rain-sprinkler-lawn example before application to Carlsmith's complex model of existential risk from power-seeking AI.

Analytically, the thesis demonstrated several capabilities that address critical governance needs. Structural analysis identified central variables, critical pathways, and influence patterns in the formalized models. Probabilistic assessment provided sensitivity analysis, uncertainty propagation, and conditional relationship exploration. Policy evaluation enabled counterfactual analysis, intervention comparison, and portfolio assessment. These capabilities enhance governance discourse by making assumptions explicit, relationships precise, and analysis systematic.

In governance terms, the thesis addressed the coordination crisis through tools that bridge technical and policy domains, enhance cross-stakeholder understanding, and support strategic coordination. The approach facilitates integration across different perspectives by creating shared representations with multiple levels of engagement, from basic causal structure to detailed probability analysis. This creates epistemic infrastructure for more productive discourse about risk factors, governance options, and intervention priorities.

The application to Carlsmith's model demonstrated these contributions in practice. The formalization successfully captured the complex causal structure and probability judgments from his paper, replicating his bottom-line estimate while revealing structural insights and sensitivity patterns. The analysis identified "Misaligned_Power_Seeking" as both structurally central and highly sensitive, confirming its critical role in the risk pathway. The policy evaluation demonstrated different governance options, with monitoring and feedback systems showing unexpectedly high impact compared to more direct interventions.

What sets this research apart from previous approaches is the automated extraction process that dramatically reduces the labor intensity of formal modeling. Where manual approaches like the original MTAIR framework required days of expert time to formalize arguments, the AMTAIR approach accomplishes this in minutes, enabling broader application to the growing volume of AI safety literature. The hybrid representation preserves narrative richness alongside mathematical precision, creating bridges between qualitative argumentation and quantitative analysis.

The research also distinguished itself through the interactive visualization system that makes complex probabilistic models accessible to diverse stakeholders. By using visual encoding for probability information, progressive disclosure for complexity management, and interactive exploration for personalized engagement, the system creates multiple entry points based on different backgrounds and interests. This accessibility enhances cross-domain communication without requiring all stakeholders to adopt specialized technical vocabulary.

Together, these contributions demonstrate a novel approach to addressing the coordination crisis in AI governance—one that leverages frontier AI technologies to enhance human coordination rather than replacing human judgment. By making implicit models explicit, enabling cross-worldview comparison, and supporting policy evaluation across diverse scenarios, the AMTAIR approach creates epistemic infrastructure for more effective coordination on what may be humanity's most consequential technological challenge.

7.2 Limitations of the Current Implementation

While the AMTAIR approach demonstrates promising capabilities, the current implementation has important limitations that constrain its immediate application and suggest directions for future work. These limitations span technical, conceptual, practical, and ethical dimensions, each affecting different aspects of the system's utility and impact.

From a technical perspective, several limitations affect extraction quality and computational performance:

First, **extraction accuracy varies across different argument types**, with better performance on well-structured causal arguments than on complex normative or conceptual discussions. The current approach works well for papers like Carlsmith's that present explicit causal structures with numerical estimates, but struggles with more implicit or qualitative arguments common in philosophical AI safety literature. The system shows lower recall for complex causal expressions where influence is described across multiple sentences or through implicit relationships.

Second, **probability estimation for conditional relationships** remains challenging, particularly for variables with multiple parents. The accuracy metrics showed lower performance for conditional probability extraction (80% F1 score) compared to structural extraction (92% F1 score for node identification), reflecting the greater complexity of quantifying relationships between variables. This limitation becomes more significant as network complexity increases and conditional relationships involve more variables.

Third, **computational scalability faces barriers for very large networks**, though current performance remains reasonable for practical applications. While the implementation handles Carlsmith's 21-node model efficiently (processing in approximately 18 seconds), much larger networks with hundreds of nodes might face computational barriers, particularly for inference operations requiring exact probability calculations. The current optimization level prioritizes accuracy over performance, with several opportunities for efficiency improvements not yet implemented.

Conceptually, the approach includes simplifications and assumptions that limit its representational capacity:

First, **temporal dynamics receive limited representation** in the current Bayesian network formalism, which captures causal structure but not explicit temporal evolution. This creates challenges for modeling dynamic processes like technological development trajectories or institutional adaptation, which might involve feedback loops or path dependencies better represented in dynamic models. The current implementation treats these dynamics implicitly through causal structure rather than modeling them explicitly.

Second, **value diversity and normative considerations** lack structured representation beyond basic variables and probabilities. While the system can represent different empirical judgments across worldviews, it provides less structure for representing different value frameworks that might influence how outcomes are evaluated. Normative considerations can be included as variables (e.g., “Stakes_Of_Error” in Carlsmith’s model), but their special status as evaluative rather than descriptive factors lacks explicit representation.

Third, **uncertainty representation focuses primarily on parameter uncertainty** rather than deeper forms of uncertainty about model structure or conceptual frameworks. While the system represents uncertainty about probability values, it provides less support for representing fundamental uncertainty about which variables matter or how they relate causally. This limitation affects how the system handles deep uncertainty characteristic of transformative AI governance.

Practically, several constraints affect real-world implementation and adoption:

First, **integration with existing governance processes** remains at a conceptual rather than operational level. The current implementation focuses on the technical pipeline without detailed workflows for incorporating the approach into specific governance contexts like technical standards development, regulatory impact assessment, or multi-stakeholder initiatives. This integration gap limits immediate practical application despite the demonstrated technical capabilities.

Second, **validation relies primarily on internal consistency** rather than extensive empirical testing against outcomes. Given the forward-looking nature of existential risk assessment, traditional validation against observed outcomes remains challenging, limiting confidence about model accuracy in representing complex real-world dynamics. The current approach emphasizes conceptual validation and expert assessment rather than empirical testing.

Third, **usability testing with diverse stakeholders** remains limited, with interface design based primarily on principles rather than extensive user research. While the visualization system incorporates accessibility features like progressive disclosure and visual encoding, these design choices haven’t been extensively validated with the diverse stakeholders who might engage with the system in governance contexts.

Ethically, several considerations affect responsible implementation and use:

First, **potential misinterpretation or overconfidence** remains a risk despite explicit uncertainty representation. Users might interpret visual models as more definitive than warranted, particularly if they focus on the intuitive visualization without engaging with uncertainty information. This risk requires careful attention to presentation and documentation to maintain appropriate epistemic humility.

Second, **accessibility barriers might affect participation** despite efforts to create multilevel interfaces. Stakeholders with limited technical backgrounds or different cultural contexts might still face challenges engaging with the formal representations, potentially creating disparities in influence over governance discussions. This concern requires ongoing attention to inclusive design and complementary engagement methods.

Third, **value-laden decisions in model construction** might remain implicit despite efforts at transparency. Choices about which variables to include, how to structure relationships, and what probabilities to assign inevitably involve value judgments that might not be fully explicit even in formalized representations. This limitation requires careful attention to documentation and transparent modeling processes.

These limitations don't fundamentally undermine the approach but highlight important areas for refinement and extension. The current implementation represents a promising foundation with clear paths for enhancement rather than a comprehensive solution to the coordination challenges in AI governance. Acknowledging these limitations demonstrates epistemic humility while suggesting concrete directions for future research and development.

7.3 Future Research Directions

The limitations identified in the previous section suggest several promising directions for future research that could enhance the AMTAIR approach and extend its applications. These directions span technical improvements, integration pathways, application domains, and theoretical extensions, each offering opportunities to build on the current foundation.

Technical enhancements could significantly improve extraction quality, analytical capabilities, and computational performance:

First, **enhanced extraction techniques** could address current limitations in handling complex arguments. Approaches might include:

- Context-aware extraction that considers document-wide information rather than isolated passages
- Multi-step reasoning that breaks complex arguments into simpler components before integration
- Comparative extraction using multiple frontier LLMs to identify areas of convergence and divergence
- Few-shot learning with expert-validated examples to improve performance on edge cases

These techniques would enhance the system's ability to handle diverse argumentation styles and complex causal expressions, expanding the range of literature it can effectively process.

Second, **advanced visualization approaches** could improve accessibility and insight generation:

- Adaptive visualization that adjusts complexity based on user background and interests
- Comparative views that highlight differences between worldviews or intervention scenarios
- Uncertainty visualization techniques that more intuitively represent different forms of uncertainty

- Temporal evolution views that show how networks change over time as new information emerges

These visualization enhancements would make complex models more accessible to diverse stakeholders while revealing patterns that might not be apparent in static representations.

Third, **improved inference algorithms** could enhance computational performance and analytical capabilities:

- Approximate inference methods for handling larger networks more efficiently
- Specialized algorithms for policy evaluation and intervention comparison
- Distributed computation for processing multiple models or scenarios in parallel
- Progressive computation that provides initial results quickly while refining with additional resources

These algorithmic improvements would enable analysis of more complex models while maintaining interactive performance for real-time exploration.

Integration pathways present opportunities to connect the AMTAIR approach with complementary systems and data sources:

First, **prediction market integration** could enable dynamic updating based on forecasting data:

- API connections to platforms like Metaculus, Manifold, and Polymarket
- Semantic mapping between forecast questions and model variables
- Automated updating of probability distributions based on forecast changes
- Relevance calculation to identify which forecasts would most reduce model uncertainty

This integration would transform static models into dynamic representations that evolve as new information emerges, addressing the current limitation of models becoming outdated quickly in rapidly changing domains.

Second, **literature monitoring systems** could automate model updates based on new research:

- Continuous scanning of AI safety literature for relevant publications
- Incremental updating of existing models rather than complete reconstruction
- Conflict detection when new findings contradict existing model assumptions
- Trend analysis to identify emerging themes and shifting consensus

This capability would help models remain current with evolving research, ensuring their ongoing relevance for governance discussions.

Third, **collaborative modeling platforms** could enable distributed development and critique:

- Multi-user interfaces for collaborative model construction and refinement
- Annotation and commenting features for model critique and discussion
- Version control for tracking model evolution over time
- Permission systems for managing contribution and review processes

These collaborative features would support community engagement with model development, enhancing both quality and legitimacy through broader participation.

Application domains beyond the current focus offer opportunities to demonstrate broader utility:

First, **other existential risk domains** might benefit from similar formalization approaches:

- Biosecurity governance for managing risks from advanced biotechnology
- Nuclear security coordination for preventing catastrophic conflicts
- Climate governance for addressing extreme climate scenarios
- Emerging technology governance beyond AI

These applications would leverage the same methodological approach while addressing different substantive domains, potentially revealing common patterns across existential risk governance challenges.

Second, **complex policy challenges** beyond existential risk might benefit from formal modeling:

- Public health policy for managing pandemic responses
- Economic policy for addressing systemic financial risks
- Environmental policy for managing ecosystem tipping points
- Technology policy for governing emerging technologies

These applications would test the approach's utility for more immediate governance challenges, potentially creating broader adoption pathways while addressing current societal needs.

Third, **organizational strategy development** presents opportunities for applied formalization:

- Research prioritization for AI safety organizations
- Grant making strategy for philanthropic funders
- Corporate risk assessment for technology companies
- Institutional design for governance bodies

These practical applications would connect formalization to concrete decision contexts, demonstrating utility beyond academic analysis to operational strategy development.

Theoretical extensions could expand the conceptual foundations and analytical capabilities:

First, **enhanced uncertainty representation** could address limitations in handling deep uncertainty:

- Multi-model ensembles to represent structural uncertainty about causal relationships
- Second-order probabilities to capture uncertainty about probability judgments
- Imprecise probabilities and interval estimates for representing ambiguity
- Non-probabilistic uncertainty representations for truly novel scenarios

These approaches would enhance the system's ability to represent the deep uncertainty characteristic of transformative technology governance, supporting more robust analysis under various forms of uncertainty.

Second, **value representation frameworks** could improve handling of normative considerations:

- Multi-attribute utility structures for representing different value priorities
- Value sensitivity analysis to show how different normative assumptions affect conclusions
- Explicit separation of empirical and normative components in models
- Comparative evaluation frameworks across different value systems

These extensions would enhance the system's ability to represent how different value frameworks influence risk assessment and intervention evaluation, making normative dimensions more explicit.

Third, **integration with formal theories** from relevant disciplines could enhance analytical foundations:

- Decision theory for modeling choice under uncertainty
- Game theory for representing strategic interactions between actors
- Institutional design theory for modeling governance structures
- Complex systems theory for understanding emergent dynamics

These theoretical integrations would strengthen the conceptual foundations of the approach while enabling more sophisticated analysis of governance challenges.

A concrete research agenda emerging from these directions might prioritize:

1. Improving extraction quality for complex arguments through context-aware techniques
2. Developing prediction market integration for dynamic model updating
3. Enhancing visualization accessibility through adaptive interfaces
4. Extending uncertainty representation to address deeper forms of uncertainty
5. Creating collaborative modeling platforms for community engagement

This agenda balances technical enhancements with practical applications, addressing key limitations while expanding potential impact. The prioritization reflects both feasibility considerations and potential value for addressing the coordination crisis in AI governance.

These future directions demonstrate the rich potential for building on the current foundation, addressing limitations while expanding capabilities and applications. The AMTAIR approach represents not a final solution but an initial step toward computational tools that enhance human coordination on complex governance challenges—a direction that becomes increasingly valuable as AI capabilities continue to advance and the window for establishing effective governance narrows.

7.4 Broader Implications for AI Governance

Beyond specific technical contributions and future research directions, this work has broader implications for how we approach AI governance challenges, particularly those related to existential risk from advanced systems. These implications touch on epistemics, coordination mechanisms, strategic planning, institutional design, and normative considerations that extend beyond the specific methodology developed in this thesis.

From an epistemic perspective, the AMTAIR approach demonstrates the value of making implicit models explicit through formalization. Much of the current discourse about AI risk involves implicit causal models and probability judgments that remain unstated or ambiguous, creating barriers to productive discourse. By providing tools that formalize these implicit models, we create foundations for more rigorous evaluation, comparison, and refinement of governance approaches.

This epistemic transformation parallels developments in other scientific domains, where formalization has enabled more rapid progress by creating shared reference points for discourse. Just as mathematical formalization accelerated physics by providing precise representations of physical theories, computational formalization of governance models can enhance progress by enabling more precise articulation and comparison of governance approaches. The AMTAIR approach contributes to this epistemic infrastructure through automated extraction and standardized representations.

For coordination mechanisms, the research highlights the importance of shared representations that bridge domain boundaries. The current coordination crisis stems partly from incompatible languages and frameworks across technical, governance, and ethical domains, creating barriers to alignment even when substantive agreement exists. By creating representations that maintain connections to multiple domains, we enable more effective coordination without requiring all stakeholders to adopt a single specialized language.

This insight suggests broader implications for governance design: effective coordination doesn't require consensus on all aspects but rather shared interfaces that enable productive interaction despite differences in background and perspective. The AMTAIR approach demonstrates one such interface through interactive visualizations that provide multiple entry points while maintaining a consistent underlying representation. This principle might inform other coordination mechanisms beyond computational tools.

In strategic planning, the formalization approach enables more systematic reasoning about intervention impacts across different scenarios. Current governance discussions often focus on abstract principles rather than concrete causal mechanisms, creating challenges for evaluating how different approaches might perform under various conditions. By modeling specific causal pathways and enabling counterfactual analysis, we create foundations for more robust strategic planning that acknowledges deep uncertainty while identifying interventions likely to perform well across scenarios.

This capability connects to growing interest in robust decision-making under deep uncertainty, which seeks approaches that work reasonably well across different possible futures rather than optimizing for specific scenarios. The policy evaluation demonstrated in this thesis supports this robust approach by enabling systematic comparison of interventions across different assumptions, helping identify governance options that maintain value under various conditions.

For institutional design, the research suggests the importance of epistemic infrastructure alongside traditional governance structures. Many current governance discussions focus on institutional forms (agencies, standards bodies, international agreements) without sufficient attention to the knowledge infrastructure needed for effective coordination. The AMTAIR approach highlights how computational tools can enhance this epistemic dimension, supporting more effective coordination regardless of specific institutional arrangements.

This perspective suggests complementary priorities for governance development: alongside formal institutions and agreements, we need investments in tools, practices, and infrastructure that support collective sense-making about complex challenges. The approach demonstrated in this thesis represents one component of such infrastructure, focusing on formal modeling while complementing other approaches like forecasting platforms, collaborative research initiatives, and cross-stakeholder dialogue processes.

From a normative perspective, the research raises important questions about values in governance design. The formalization approach doesn't eliminate value judgments but rather makes them more explicit through model construction choices and parameter settings. This explicitness creates opportunities for more transparent discourse about how different value frameworks influence risk assessment and intervention evaluation, potentially leading to governance approaches that better reflect diverse perspectives.

This normative dimension connects to broader questions about inclusivity and legitimacy in AI governance. As the field develops, ensuring representation of diverse perspectives becomes increasingly important for both technical quality and moral legitimacy. The accessibility features of the AMTAIR approach represent initial steps toward more inclusive formalization, but much more work remains to ensure governance processes incorporate appropriate diversity of perspective.

Looking ahead, the accelerating pace of AI capability development creates urgency for effective governance coordination. Recent advances in frontier models demonstrate capabilities emerging faster than many expected, compressing available response time for governance development. This acceleration highlights the value of tools that enhance coordination efficiency, helping diverse stakeholders align efforts more rapidly than traditional processes might allow.

The AMTAIR approach contributes to this coordination challenge by creating computational infrastructure that makes implicit models explicit, facilitates cross-domain communication, and enables systematic evaluation of governance options. While technical tools alone cannot solve the coordination crisis, they can enhance human coordination capabilities precisely when such enhancement becomes most necessary.

In conclusion, this research demonstrates that computational approaches to formalization, when thoughtfully designed with appropriate attention to accessibility and uncertainty representation, can enhance rather than undermine human coordination on complex governance challenges. By creating bridges between qualitative argumentation and quantitative analysis, making implicit models explicit, and enabling systematic comparison across perspectives, such approaches provide valuable infrastructure for addressing what may be humanity's most consequential technological challenge.

The path forward involves not just technical development but thoughtful integration with broader governance processes, combining computational tools with human judgment, institutional design, and normative reflection. The AMTAIR approach represents an initial step in this direction, with promising potential for enhancing our collective ability to govern advanced AI systems wisely and effectively.

8. References

[Note: This section would contain a comprehensive bibliography organized by topic area, including primary sources for AI safety and governance literature, technical references for Bayesian networks and computational methods, sources for the Carlsmith model and other risk assessments, and methodological references for formal modeling in governance contexts.]

Appendix A: Technical Implementation Details

[Note: This appendix would provide detailed technical information about the implementation, including environment setup instructions, full code listings for key components, API specifications, data format definitions, and documentation of the development workflow. This material supports reproducibility and extension of the research.]

Appendix B: BayesDown Syntax Specification

[Note: This appendix would provide a comprehensive specification of the BayesDown syntax, including formal grammar definitions, validation rules, extension mechanisms, and guidelines for converting between different representation formats. This material enables other researchers to use and extend the intermediate representation format developed in this research.]

Appendix C: Complete Carlsmith Model Analysis

[Note: This appendix would include the complete formalized representation of Carlsmith's model, detailed explanation of how probabilities were derived from his text, comprehensive analysis results, discussion of alternative interpretations, and documentation of validation with domain experts. This material provides a complete case study demonstrating the AMTAIR approach applied to a complex real-world risk assessment.]

Appendix D: Additional Case Studies

[Note: This appendix would present additional applications of the AMTAIR approach to other AI risk frameworks, real-world policy scenarios, and comparative analyses with manual approaches. This material demonstrates broader applicability beyond the primary case study examined in the main text.]