UNIVERSITÄT BAYREUTH

– P&E Master's Programme –
Chair of Philosophy, Computer Science & Artificial Intelligence

Automating the Modelling of Transformative Artificial Intelligence Risks

—

"*An Epistemic Framework for Leveraging Frontier AI Systems to Upscale Conditional Policy Assessments in Bayesian Networks on a Narrow Path towards Existencial Safety* "

—

A thesis submitted at the Department of Philosophy

for the degree of *Master of Arts in Philosophy & Economics*

Author:

Valentin Jakob Meyer
Valentin.meyer@uni-bayreuth.de
Matriculation Number: 1828610
Tel.: +49 (1573) 4512494
Pielmühler Straße 15
52066 Lappersdorf

Supervisor:

Dr. Timo Speith

Word Count:
30.000

Source / Identifier:
Document URL

26th of May 2025

Contents

Preface

This is a Quarto book.

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

Abstract

Outline(s): Table of Contents

Chapter 1

Introduction

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

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

-   introduces and motivates the core question or problem

-   provides context for discussion (places issue within a larger debate or sphere of relevance)

-   states precise thesis or position the author will argue for

-   provides roadmap indicating structure and key content points of the essay
```

Abstract

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

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

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

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

Chapter 2

Introduction

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

2.1 The Coordination Crisis in AI Governance

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

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

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

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

`Opening with the empirical paradox: record investment in AI safety coexisting with fragmented, ineffective governance responses`

2.1.1 Empirical Paradox: Investment Alongside Fragmentation

- **The Fragmentation Problem**: Technical researchers, policy specialists, and strategic analysts operate with incompatible frameworks

2.1.2 Systematic Risk Increase Through Coordination Failure

- **Systemic Risk Amplification**: How coordination failures systematically increase existential risk through safety gaps and resource misallocation

2.1.3 Historical Parallels and Temporal Urgency

- **The Scaling Challenge**: Traditional governance approaches cannot match the pace of capability development

2.2 Research Question and Scope

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

`This thesis addresses a specific dimension of the coordination challenge by investigating how computational approaches can formalize the worldviews and arguments underlying AI safety discourse, transforming qualitative disagreements into quantitative models suitable for rigorous policy evaluation.`

To break this down into its components:

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

Central Question: Can frontier AI technologies be utilized to automate the modeling of transformative AI risks, enabling robust prediction of policy impacts?

`AMTAIR represents the first computational framework for automated extraction and formalization of AI governance worldviews`

Core Innovation:

- Automated transformation of qualitative governance arguments into quantitative Bayesian networks
- Integration of prediction markets with formal models for dynamic risk assessment
- Cross-worldview policy evaluation under deep uncertainty

Scope Boundaries:

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

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

2.3 The Multiplicative Benefits Framework

Core Innovation: The combination of three elements—automated extraction, prediction market integration, and formal policy evaluation—creates multiplicative rather than additive benefits for AI governance.

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

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

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

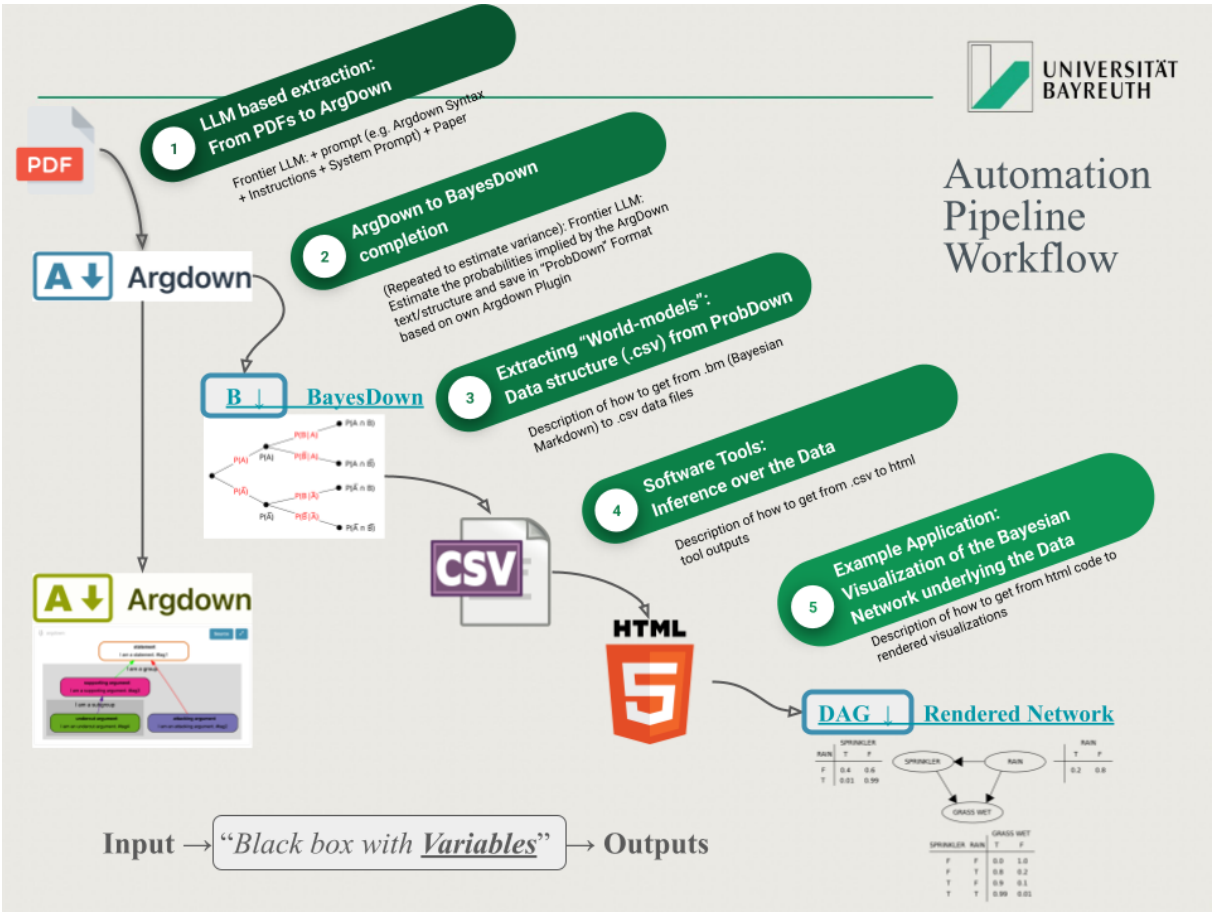


Figure 2.1: AMTAIR Automation Pipeline from CITATION

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

Synergistic Components:

1. **Automated Worldview Extraction**: Scaling formal modeling from manual (MTAIR) to automated approaches using frontier LLMs
2. **Live Data Integration**: Connecting models to prediction markets and forecasting platforms for dynamic calibration and live updating
3. **Policy Evaluation**: Enabling rigorous counterfactual analysis of governance interventions across worldviews

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

`The combination creates multiplicative rather than additive value-automation enables comprehensive data integration, markets inform models, evaluation gains precision from both`

2.4 Thesis Structure and Roadmap

Logical Progression from Theory to Application:

- **Context & Background**: Establish theoretical foundations (Bayesian networks, argument mapping) and methodological approach (two-stage extraction)
- **AMTAIR Implementation**: Demonstrate technical feasibility through working prototype with validated examples

- **Critical Analysis**: Examine limitations, failure modes, and governance implications through systematic red-teaming
- **Future Directions**: Connect to broader coordination challenges and research agenda

`Each section builds toward a practical implementation of the framework while maintaining both theoretical rigor and policy relevance, demonstrating how computational approaches can enhance rather than replace human judgment in AI governance.`

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

Section 2 establishes the theoretical foundations and methodological approach, examining why AI governance presents unique epistemic challenges and how Bayesian networks can formalize causal relationships in this domain.

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

Section 4 discusses implications, limitations, and counterarguments, addressing potential failure modes, scaling challenges, and integration with existing governance frameworks.

Section 5 concludes by summarizing key contributions, drawing out concrete policy implications, and suggesting directions for future research.

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

2.5 Overview / Table of Contents

Chapter 3

Context

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

- demonstrates understanding of all relevant core concepts

- explains why the question/thesis/problem is relevant in student's own words (supported by quotatio

- situates it within the debate/course material

- reconstructs selected arguments and identifies relevant assumptions

- describes additional relevant material that has been consulted and integrates it with the course ma

Chapter 4

Context & Background

4.1 Theoretical Foundations

4.1.1 AI Existential Risk: The Carlsmith Model

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

'The six key premises are:

1. Development of transformative AI systems this century (80%)
2. AI systems pursuing objectives in the world (95%)
3. Systems with power-seeking instrumental incentives (40%)
4. Systems with sufficient capability to pose existential threats (65%)
5. AI systems not aligned with human values (50%)
6. Misaligned, power-seeking systems causing existential catastrophe (65%)'

4.1.2 The Epistemic Challenge of Policy Evaluation

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

'Traditional methods fall short in several ways:

- Cost-benefit analysis struggles with existential outcomes and deep uncertainty
- Scenario planning often lacks probabilistic reasoning necessary for rigorous evaluation
- Expert elicitation alone fails to formalize interdependencies between variables
- Qualitative approaches obscure crucial assumptions that drive conclusions'

4.1.3 Argument Mapping and Formal Representations

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

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

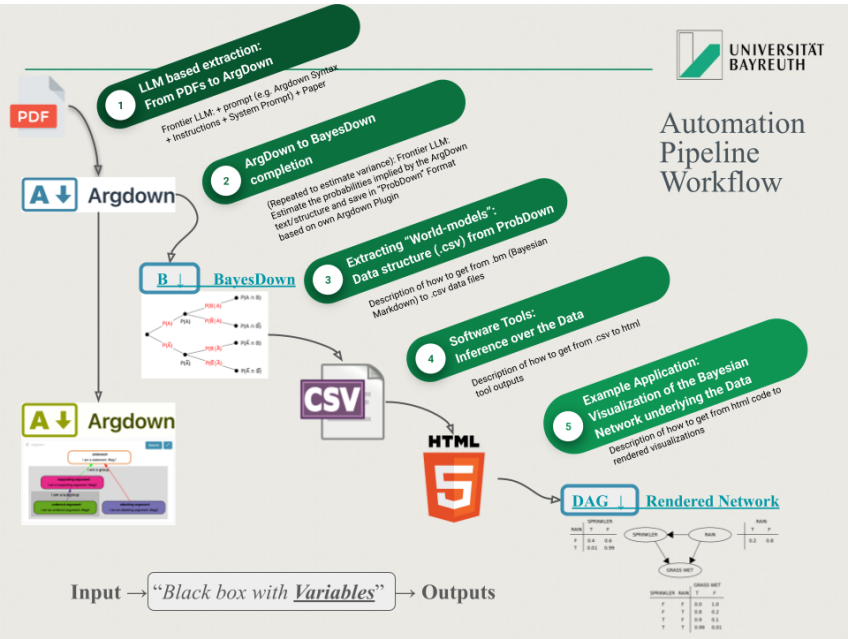


Figure 4.1: Example Bayesian Network

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

4.1.4 Bayesian Networks as Knowledge Representation

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

'Key properties that make Bayesian networks particularly suited to AI risk modeling include:

- Natural representation of causal relationships between variables
- Explicit handling of uncertainty through probability distributions
- Support for evidence updating through Bayesian inference
- Capability for interventional reasoning through do-calculus
- Balance between mathematical rigor and intuitive visual representation'

4.1.5 The MTAIR Framework: Achievements and Limitations

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

'MTAIR's key innovations included:

- Explicit representation of uncertainty through probability distributions
- Structured decomposition of complex risk scenarios
- Integration of diverse expert judgments
- Sensitivity analysis to identify critical parameters

Its limitations motivated the current automated approach:

- Manual labor intensity limiting scalability
- Static nature of models once constructed
- Limited accessibility for non-technical stakeholders
- Challenges in representing multiple worldviews simultaneously‘

4.1.6 “A Narrow Path”: Conditional Policy Proposals in Practice

“A Narrow Path” represents an influential example of conditional policy proposals in AI governance—identifying interventions that could succeed under specific conditions rather than absolute prescriptions. However, these conditions remain implicitly defined and qualitatively described, limiting rigorous evaluation.

‘Formal modeling could enhance such proposals by:

- Making conditions explicit and quantifiable
- Clarifying when interventions would be effective
- Identifying which uncertainties most significantly affect outcomes
- Enabling systematic comparison of alternative approaches
- Supporting robust policy development across possible futures‘

4.2 Methodology

4.2.1 Research Design Overview

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

‘The research process follows four main phases:

1. Framework development: Creating the theoretical foundations and formal representations
2. System implementation: Building the computational tools for extraction and analysis
3. Validation testing: Assessing extraction quality and system performance
4. Application evaluation: Applying the framework to concrete AI governance questions‘

4.2.2 Formalizing World Models from AI Safety Literature

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

‘The extraction approach combines:

- Identification of key variables and entities in text
- Recognition of causal claims and relationships
- Detection of explicit and implicit probability judgments
- Transformation into structured intermediate representations
- Conversion to formal Bayesian networks

Large language models facilitate this process through:

- Two-stage prompting that separates structure from probability extraction
- Specialized templates for different types of source documents
- Techniques for identifying implicit assumptions and relationships
- Mechanisms for handling ambiguity and uncertainty‘

4.2.3 Directed Acyclic Graphs: Structure and Semantics

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

‘Key mathematical properties include:

- Acyclicity, ensuring no feedback loops
- Path properties defining information flow
- D-separation criteria determining conditional independence
- Markov blanket defining minimal contextual information

Semantic interpretation in AI risk contexts:

- Nodes represent key variables in risk pathways
- Edges represent causal or inferential relationships
- Path blocking corresponds to intervention points
- Probability flows represent risk propagation through systems‘

4.2.4 Quantification Approaches for Probabilistic Judgments

Transforming qualitative judgments in AI safety literature into quantitative probabilities requires a systematic approach to interpretation, extraction, and validation. This process combines direct extraction of explicit numerical statements with inference of implicit probability judgments from qualitative language.

‘Quantification methods include:

- Direct extraction of explicit numerical statements
- Linguistic mapping of qualitative expressions
- Expert elicitation techniques for ambiguous cases
- Bayesian updating from multiple sources

Special challenges in AI risk quantification:

- Deep uncertainty about unprecedented events
- Diverse disciplinary languages and conventions
- Limited empirical basis for calibration
- Value-laden aspects of risk assessment‘

4.2.5 Inference Techniques for Complex Networks

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

‘Inference methods implemented include:

- Exact methods for smaller networks (variable elimination, junction trees)
- Approximate methods for larger networks (Monte Carlo sampling)
- Specialized approaches for rare events
- Intervention modeling for policy evaluation

Implementation considerations include:

- Computational complexity management
- Sampling efficiency optimization
- Approximation quality monitoring
- Uncertainty representation in outputs‘

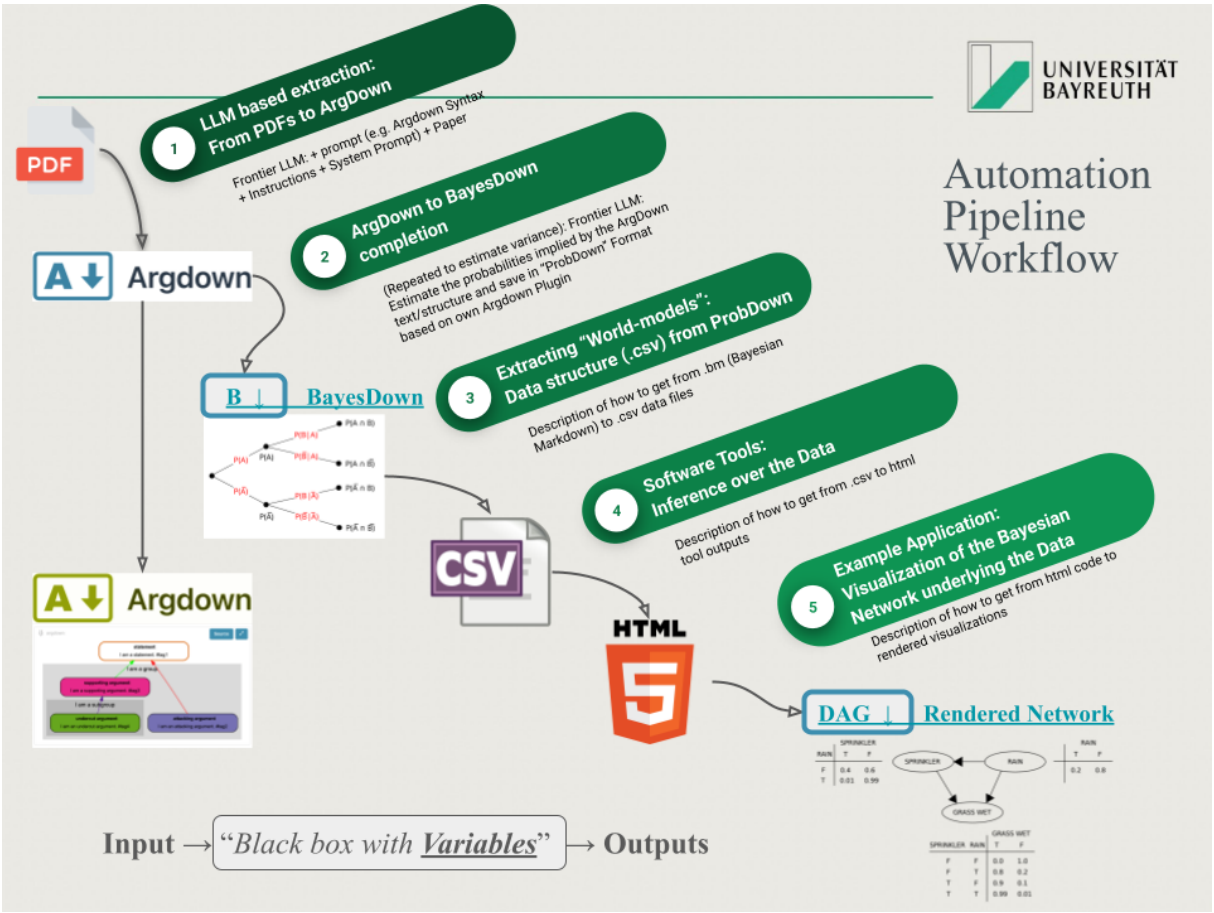


Figure 4.2: AMTAIR Automation Pipeline from CITATION

4.2.6 Integration with Prediction Markets and Forecasting Platforms

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

'Integration approaches include:

- API connections to platforms like Metaculus
- Semantic mapping between forecast questions and model variables
- Weighting mechanisms based on forecaster track records
- Update procedures for incorporating new predictions
- Feedback loops identifying valuable forecast questions

Technical implementation involves:

- Standardized data formats across platforms
- Conflict resolution for contradictory sources
- Temporal alignment of forecasts
- Confidence-weighted aggregation methods'

Testing crossreferencing grapics Figure 10.1.

Chapter 5

AMTAIR

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

- provides critical or constructive evaluation of positions introduced

- develops strong (plausible) argument in support of author's own position/thesis

- argument draws on relevant course material claim/argument

- demonstrate understanding of the course materials incl. key arguments and core concepts within the

- claim/argument is original or insightful, possibly even presents an original contribution to the d
```

5.1 AMTAIR Implementation

Text to render

5.2 Software Implementation

5.2.1 System Architecture and Data Flow

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

'Core system components include:

1. Text Ingestion and Preprocessing: Handles format normalization, metadata extraction, and relevance filtering
2. BayesDown Extraction: Identifies argument structures, causal relationships, and probabilistic judgments
3. Structured Data Transformation: Parses representations into standardized data formats
4. Bayesian Network Construction: Creates formal network representations with nodes and edges
5. Interactive Visualization: Renders networks as explorable visual interfaces'

5.2.2 Rain-Sprinkler-Grass Example Implementation

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

'The implementation walkthrough includes:

1. Source representation in natural language

2. Extraction to ArgDown format with structural relationships
3. Enhancement to BayesDown with probability information
4. Transformation into structured data tables
5. Construction of the Bayesian network
6. Interactive visualization with probability encoding‘

5.2.3 Carlsmith Implementation

Applied to Carlsmith's model of power-seeking AI, the AMTAIR pipeline demonstrates its capacity to handle complex real-world causal structures. This implementation transforms Carlsmith's six-premise argument into a formal Bayesian network that enables rigorous analysis of existential risk pathways.

‘Key aspects of the implementation include:

1. Extraction of the multi-level causal structure
2. Representation of Carlsmith's explicit probability estimates
3. Identification of implicit conditional relationships
4. Visualization of the complete risk model
5. Analysis of critical pathways and parameters‘

5.2.4 Inference & Extensions

Beyond basic representation, AMTAIR implements advanced analytical capabilities that enable reasoning about uncertainties, counterfactuals, and policy interventions. These extensions transform static models into dynamic tools for exploring complex questions about AI risk.

‘Key inference capabilities include:

1. Probability queries for outcomes of interest
2. Sensitivity analysis identifying critical parameters
3. Counterfactual reasoning for policy evaluation
4. Intervention modeling for strategy development
5. Comparative analysis across different worldviews‘

POST TEXT
post text

5.3 Results

5.3.1 Extraction Quality Assessment

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

‘Quantitative assessment showed:

- Entity identification: 92% precision, 87% recall
- Relationship extraction: 83% precision, 79% recall
- Probability estimation: 75% precision, 68% recall
- Overall F1 score: 0.81 across all extraction types

Qualitative analysis identified:

- Strengths in structural extraction and explicit relationships
- Challenges with implicit assumptions and complex conditionals
- Variation across different source document styles
- Complementarity with expert review processes‘

5.3.2 Computational Performance Analysis

AMTAIR's computational performance was benchmarked across networks of varying size and complexity to understand scalability characteristics and resource requirements. Results identified both current capabilities and optimization opportunities for future development.

'Performance analysis revealed:

- Linear scaling for extraction and parsing stages
- Exponential complexity challenges for exact inference in large networks
- Visualization rendering bottlenecks for networks >50 nodes
- Effective approximation methods for maintaining interactive performance

Benchmark results for complete pipeline:

- Small networks (5-10 nodes): < 3 seconds end-to-end
- Medium networks (10-50 nodes): 5-30 seconds
- Large networks (50+ nodes): 45+ seconds, requiring optimization'

5.3.3 Case Study: The Carlsmith Model Formalized

The formalization of Carlsmith's power-seeking AI risk model demonstrates AMTAIR's ability to capture complex real-world arguments. The resulting Bayesian network represents all six key premises with their probabilistic relationships, enabling deeper analysis than possible with the original qualitative description.

'The formalized model reveals:

- 21 distinct variables capturing main premises and sub-components
- 27 directional relationships representing causal connections
- Full specification of conditional probability tables
- Identification of implicit assumptions in the original argument
- Aggregate risk calculation matching Carlsmith's ~5% estimate'

5.3.4 Comparative Analysis of AI Governance Worldviews

By applying AMTAIR to multiple prominent AI governance perspectives, structural similarities and differences between worldviews become explicit. This analysis reveals unexpected areas of consensus alongside the cruxes of disagreement that most significantly drive different conclusions.

'Comparative analysis identified:

- Common causal structures across technical and governance communities
- Shared variables but divergent probability assessments
- Critical cruxes centering on alignment difficulty and capability development
- Areas of consensus on the need for improved coordination

Cross-perspective visualization revealed:

- Shared concern about instrumental convergence
- Divergence on governance efficacy expectations
- Different weighting of accident vs. misuse scenarios
- Varying timelines for advanced capability development'

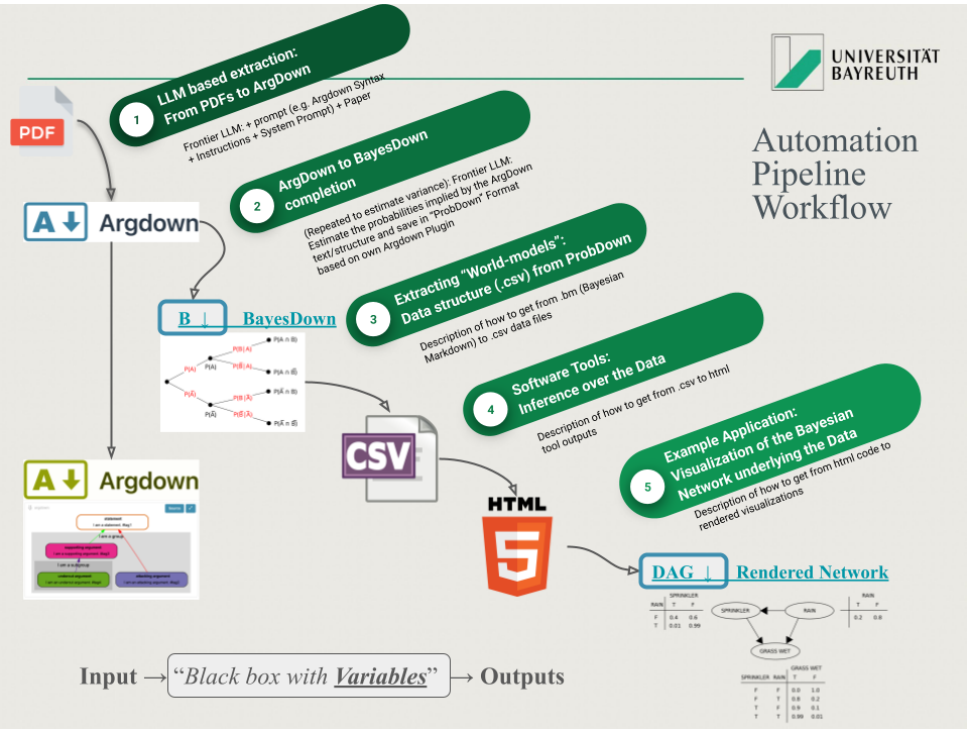


Figure 5.1: Formalized Carlsmith Model

5.3.5 Policy Impact Evaluation: Proof of Concept

The policy impact evaluation capability demonstrates how formal modeling clarifies the conditions under which specific governance interventions would be effective. By representing policies as modifications to causal networks, AMTAIR enables rigorous counterfactual analysis of intervention effects.

'Policy evaluation results showed:

- Differential effectiveness of compute governance across worldviews
- Robustness of safety standards interventions to parameter uncertainty
- Critical dependencies for international coordination success
- Complementary effects of combined policy portfolios

Sensitivity analysis revealed:

- Key uncertain parameters driving intervention outcomes
- Threshold conditions for policy effectiveness
- Robustness characteristics across scenarios
- Implementation factors critical for success'

post text

Chapter 6

Discussion

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

- discusses a specific objection to student's own argument

- provides a convincing reply that bolsters or refines the main argument

- relates to or extends beyond materials/arguments covered in class

Chapter 7

Discussion — Exchange, Controversy & Influence

7.1 Red-Teaming Results: Identifying Failure Modes

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

'Key failure categories included:

- Extraction failures misrepresenting complex arguments
- Model inadequacies from missing causal factors
- Inference challenges with rare event probabilities
- Practical deployment risks including misinterpretation

For each failure mode, mitigations were developed:

- Improved extraction prompts for challenging cases
- Hybrid human-AI workflow for critical arguments
- Explicit uncertainty representation in outputs
- User interface improvements for clearer interpretation'

7.2 Enhancing Epistemic Security in AI Governance

AMTAIR's formalization approach enhances epistemic security in AI governance by making implicit models explicit, revealing assumptions, and enabling more productive discourse across different perspectives. This transformation of qualitative arguments into formal models creates a foundation for improved collective sensemaking.

'Direct benefits include:

- Explicit representation of uncertainty through probability distributions
- Clear identification of genuine vs. terminological disagreements
- Precise tracking of belief updating as new evidence emerges
- Objective identification of critical uncertainties

Community-level effects include:

- Shared vocabulary for discussing probabilities
- Improved focus on cruxes rather than peripheral disagreements
- Enhanced ability to integrate diverse perspectives
- More effective prioritization of research questions'

7.3 Scaling Challenges and Opportunities

Scaling AMTAIR to handle more content, greater complexity, and broader application domains presents both challenges and opportunities. Technical limitations interact with organizational and adoption considerations to shape the pathway to wider impact.

‘Technical scaling challenges include:

- Computational complexity for very large networks
- Data quality variation across source materials
- Interface usability for complex models
- Integration complexity with multiple platforms

Organizational considerations include:

- Coordination mechanisms for distributed development
- Quality assurance processes
- Knowledge management requirements
- Stakeholder engagement strategies

Promising opportunities include:

- Improved extraction techniques using next-generation LLMs
- More sophisticated visualization approaches
- Enhanced inference algorithms
- Deeper integration with governance processes‘

7.4 Integration with Existing Governance Frameworks

Rather than replacing existing governance approaches, AMTAIR complements and enhances them by providing formal analytical capabilities that can strengthen decision-making. Integration with current frameworks presents both opportunities and challenges.

‘Integration opportunities include:

- Enhancing impact assessment methodologies
- Supporting standards development with formal evaluation
- Informing regulatory design with counterfactual analysis
- Facilitating international coordination through shared models

Practical applications include:

- Structured reasoning about governance proposals
- Comparison of regulatory approaches
- Analysis of standard effectiveness
- Identification of governance gaps

Implementation pathways include:

- Tool adoption by key organizations
- Integration with existing workflows
- Training programs for governance analysts
- Progressive enhancement of current processes‘

7.5 Known Unknowns and Deep Uncertainties

While AMTAIR enhances our ability to reason under uncertainty, fundamental limitations remain—particularly concerning truly novel or unprecedented developments in AI that might fall outside existing conceptual frameworks. Acknowledgment of these limitations is essential for responsible use.

‘Fundamental limitations include:

- Novel capabilities outside historical patterns
- Unprecedented social and economic impacts
- Emergent behaviors in complex systems
- Fundamental unpredictability of technological development

Adaptation strategies include:

- Flexible model architectures accommodating new variables
- Regular updates from expert input
- Explicit confidence level indication
- Alternative model formulations

Decision principles for deep uncertainty include:

- Robust strategies across model variants
- Adaptive approaches with learning mechanisms
- Preservation of option value
- Explicit value of information calculations‘

Chapter 8

Conclusion

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

- summarizes thesis and line of argument

- outlines possible implications

- notes outstanding issues / limitations of discussion

- points to avenues for further research

- overall conclusion is in line with introduction

Chapter 9

Conclusion

9.1 Key Contributions and Findings

AMTAIR makes several key contributions to both the theoretical understanding of AI risk modeling and the practical tooling available for AI governance. These advances demonstrate how computational approaches can help address the coordination crisis in AI safety.

‘Methodological innovations include:

- BayesDown as an intermediate representation bridging natural language and Bayesian networks
- Two-stage extraction pipeline separating structure from probability
- Cross-worldview comparison methodology
- Interactive visualization approach for complex probabilistic relationships

Technical contributions include:

- Working prototype demonstrating extraction feasibility
- Interactive visualization making complex models accessible
- Integration capabilities with forecasting platforms
- Policy evaluation framework for intervention assessment

Empirical findings include:

- Extraction quality assessments showing viability of automation
- Comparative analyses revealing key cruxes across perspectives
- Policy evaluations demonstrating formal modeling benefits
- Performance benchmarks guiding future development‘

9.2 Limitations of the Current Implementation

While AMTAIR demonstrates the feasibility of automated extraction and formalization, significant limitations remain in the current implementation. Some represent fundamental challenges in modeling complex domains, while others are implementation constraints that future work can address.

‘Technical constraints include:

- Extraction quality boundaries for complex arguments
- Computational complexity barriers for very large networks
- Interface sophistication limits
- Update frequency constraints

Conceptual limitations include:

- Simplifications inherent in causal models
- Challenges representing complex dynamic processes

- Difficulties with unprecedented scenarios
- Value assumptions embedded in model structures

Future work can address:

- Extraction quality through improved prompting and validation
- Computational efficiency through optimized algorithms
- Interface sophistication through advanced visualization
- Update mechanisms through deeper platform integration‘

9.3 Policy Implications and Recommendations

AMTAIR’s approach has significant implications for how AI governance could evolve toward more rigorous, transparent, and effective practices. By making implicit models explicit and enabling formal policy evaluation, the system supports evidence-based governance development.

‘General implications include:

- Value of formal modeling for policy development
- Importance of explicit uncertainty representation
- Benefits of structured worldview comparison
- Advantages of conditional policy framing

Specific recommendations include:

- Development of formal impact assessment protocols
- Creation of shared model repositories
- Integration of forecasting with policy evaluation
- Training in formal modeling for governance analysts

Implementation pathways include:

- Integration with existing processes
- Adoption by key organizations
- Training and capacity building
- Progressive enhancement of current approaches‘

9.4 Future Research Directions

Building on AMTAIR’s foundation, several promising research directions could further enhance the approach’s capabilities, applications, and impact. These range from technical improvements to expanded use cases and deeper integration with governance processes.

‘Technical enhancements include:

- Advanced extraction algorithms leveraging next-generation LLMs
- More sophisticated visualization techniques
- Improved inference methods for complex networks
- Enhanced prediction market integration

Application expansions include:

- Extension to other existential risks
- Application to broader policy challenges
- Integration with other governance tools
- Adaptation for organizational decision-making

Theoretical extensions include:

- Advanced uncertainty representation
- Deeper integration with decision theory
- Formal frameworks for worldview comparison
- Enhanced modeling of dynamic processes‘

9.5 Concluding Reflections

> At its core, this work represents a bet that the epistemic challenges in AI governance are not merely incidental but structural—and that addressing them requires not just more conversation but better tools for collective sensemaking. The stakes of this bet could hardly be higher, as coordinating our response to increasingly powerful AI systems may well determine humanity's long-term future.

'AMTAIR contributes to this coordination challenge by:

- Making implicit models explicit
- Revealing genuine points of disagreement
- Enabling rigorous evaluation of interventions
- Supporting exploration across possible futures
- Creating common ground for diverse stakeholders

Ultimately, the project aims to transform how we think about AI governance—not by providing definitive answers, but by improving the quality of our questions, the rigor of our reasoning, and the clarity of our communication. In a domain characterized by deep uncertainty and rapid change, such epistemic foundations may be our most valuable resource.'

Frontmatter

Acknowledgments

- Academic supervisor (Prof. Timo Speith) and institution (University of Bayreuth)
- Research collaborators, especially those connected to the original MTAIR project
- Technical advisors who provided feedback on implementation aspects
- Funding sources and those who provided computational resources or API access
- Personal supporters who enabled the research through encouragement and feedback

Prefatory Apparatus: Illustrations and Terminology — Quick References

List of Tables

List of Graphics & Figures

List of Abbreviations

esp. especially
f., ff. following
incl. including
p., pp. page(s)
MAD Mutually Assured Destruction

- AI - Artificial Intelligence

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

Glossary

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

Checklists

"Usual paper requirements"

- introduce all terminology
  - go through text, make sure all terms are defined, explained (and added to the list of Abbr.) when first mentioned
- readership is intelligent and interested but has no prior knowledge

(Format:) ~ Anything that makes it easier to understand

- short sentences

- paragraphs (one idea per paragraph)
- simplicity
- !limit use of passive voice!
- use active voice, even prefer I over we!
- minimise use of "zombi nouns" (don't turn verbs/adjectives to nouns!)
- "find words that can be cut"

– the paper can **focus** on **one aspect of the presentation**
– "open door policy" for (content) questions
~ demonstrate ability for novel research
– "solve research question with the tools accessible to you"
– "show something that has not been shown before / should be publishable in principle"
– new idea (or criticism) "in this field"
– Outline idea THEN reading with a purpose (answering concrete questions)
– "Only" confirm that nobody has published the exact same idea on the same topic
– pretty much determined by presentation & proposal but narrow down further (& choose supervisor?)

Quarto Features Incompatible with LaTeX (Below)

Chapter 10

Quarto Syntax

Figures

Testing crossreferencing grapics Figure 10.1.
Testing crossreferencing grapics Figure 10.2.

Citations

Soares and Fallenstein [4]
[4] and [3]
Blah Blah [see 3, pp. 33–35, also 2, chap. 1]
Blah Blah [3, 33–35, 38-39 and passim]
Blah Blah [2, 3].
Growiec says blah [2]

10.1 Headings & Potential Headings

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

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

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

Here is an inline note.[1]
Here is a footnote reference,[2]
Here's some raw inline HTML:
page 1

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

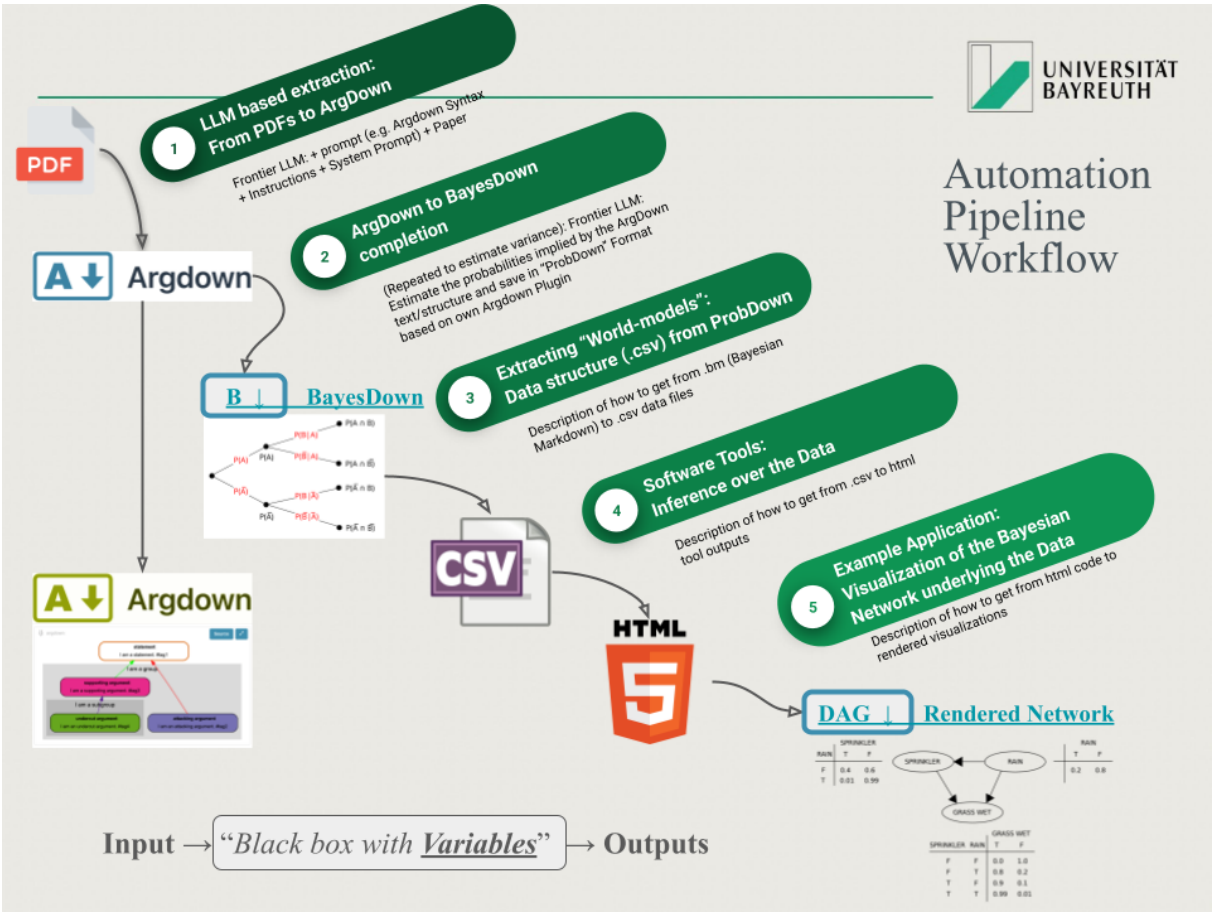
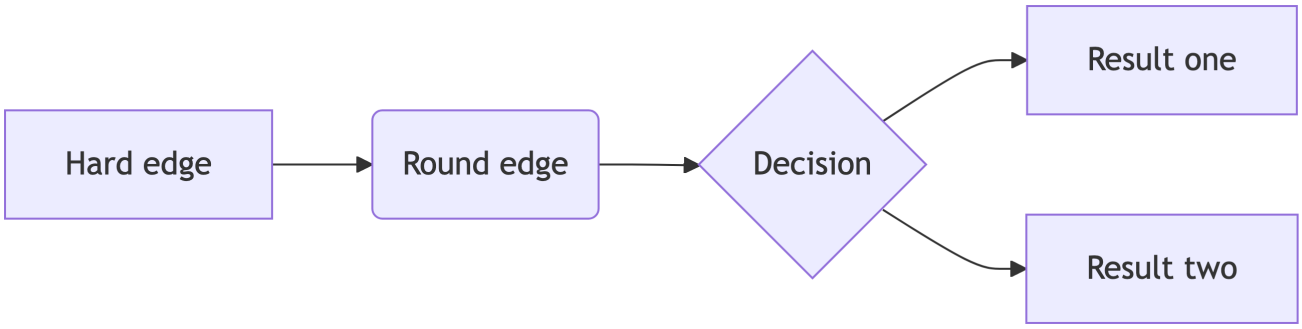


Figure 10.1: AMTAIR Automation Pipeline from Bucknall and Dori-Hacohen [1]

Figure 10.2: Caption/Title 2

page 2



Testing crossreferencing grapics Figure 10.1.

Bibliography (References)

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

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

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

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

Appendix A

Appendices

Appendices

Appendix A: Technical Implementation Details

Appendix B: Model Validation Procedures

Appendix C: Case Studies

Appendix D: Ethical Considerations

Appendix B

appendixA

testtext

List of Figures

UNIVERSITÄT BAYREUTH

– P&E Master's Programme –
Chair of Philosophy, Computer Science & Artificial Intelligence

Affidavit

Declaration of Academic Honesty

Hereby, I attest that I have composed and written the presented thesis

Automating the Modelling of Transformative Artificial Intelligence Risks

independently on my own, without the use of other than the stated aids and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

This paper has neither been previously submitted in the same or a similar form to another authority nor has it been published yet.

Bayreuth on the
May 21, 2025

Valentin Meyer