UNIVERSITÄT BAYREUTH

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

Automating the Modelling of Transformative Artificial Intelligence Risks

—

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

—

A thesis submitted at the Department of Philosophy

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

Author:

Valentin Jakob Meyer
Valentin.meyer@uni-bayreuth.de
Matriculation Number: 1828610
Tel.: +49 (1573) 4512494
Pielmühler Straße 15
93138 Lappersdorf

Supervisor:

Dr. Timo Speith

Word Count:
30.000
Source / Identifier:
Document URL

26th of May 2025

Table of Contents

List of Figures

List of Tables

Preface

Abstract

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

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

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

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

Prefatory Apparatus: Frontmatter

Illustrations and Terminology — Quick References

Acknowledgments

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

List of Graphics & Figures

List of Abbreviations

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

Glossary

- **Argument mapping**: A method for visually representing the structure of arguments
- **BayesDown**: An extension of ArgDown that incorporates probabilistic information

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

Final Thesis: Automating the Modeling of Transformative Artificial Intelligence Risks

Frontmatter: Preface

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

The journey from initial concept to working implementation involved iterative refinement based on feedback from advisors, domain experts, and potential users. What began as a technical exercise in automated extraction evolved into a broader framework for enhancing epistemic security in one of humanity's most critical coordination challenges. The AMTAIR project—Automating Transformative AI Risk Modeling—represents an attempt to build computational bridges between communities that, despite shared concerns about AI risk, often struggle to communicate effectively due to incompatible frameworks, terminologies, and implicit assumptions.

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

Acknowledgments

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

I acknowledge Johannes Meyer and Jelena Meyer for their invaluable assistance in verifying the automated extraction procedure through manual extraction of ArgDown and BayesDown data

from the Carlsmith paper, providing crucial ground truth for validation.

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

Any errors or limitations remain my own responsibility.

List of Figures

List of Tables

List of Abbreviations

AI - Artificial Intelligence
AGI - Artificial General Intelligence
AMTAIR - Automating Transformative AI Risk Modeling
API - Application Programming Interface
APS - Advanced, Planning, Strategic (AI systems)
BN - Bayesian Network
CPT - Conditional Probability Table
DAG - Directed Acyclic Graph
LLM - Large Language Model
ML - Machine Learning
MTAIR - Modeling Transformative AI Risks
NLP - Natural Language Processing
P&E - Philosophy & Economics
PDF - Portable Document Format
TAI - Transformative Artificial Intelligence

1. Introduction: The Coordination Crisis in AI Governance

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

1.1 Opening Scenario: The Policymaker's Dilemma

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

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

This scenario[1] plays out daily across government offices, corporate boardrooms, and research institutions worldwide. It exemplifies what I term the "coordination crisis" in AI governance: despite unprecedented attention and resources directed toward AI safety, we lack the epistemic infrastructure to synthesize diverse expert knowledge into actionable governance strategies **todd2024**.

Show Image

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

1.2 The Coordination Crisis in AI Governance

As AI capabilities advance at an accelerating pace—demonstrated by the rapid progression from GPT-3 to GPT-4, Claude, and emerging multimodal systems **maslej2025 samborska2025**—humanity faces a governance challenge unlike any in history. The task of ensuring increasingly powerful AI systems remain aligned with human values and beneficial to our long-term flourishing grows more urgent with each capability breakthrough. This challenge becomes particularly acute when considering transformative AI systems that could drastically alter civilization's trajectory, potentially including existential risks from misaligned systems pursuing objectives counter to human welfare.

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

The current state of AI governance presents a striking paradox. On one hand, we witness extraordinary mobilization: billions in research funding, proliferating safety initiatives, major tech companies establishing alignment teams, and governments worldwide developing AI strategies. The Asilomar AI Principles garnered thousands of signatures **tegmark2024**, the EU advances comprehensive AI regulation **european2024**, and technical researchers produce increasingly sophisticated work on alignment, interpretability, and robustness.

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

Show Image

1.2.1 Safety Gaps from Misaligned Efforts

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

When different communities operate with incompatible frameworks, critical risks fall through the cracks. Technical researchers may solve alignment problems under assumptions that policymakers' decisions invalidate. Regulations optimized for current systems may inadvertently incentivize dangerous development patterns. Without shared models of the risk landscape, our

collective efforts resemble the parable of blind men describing an elephant—each accurate within their domain but missing the complete picture **paul2023**.

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

1.2.2 Resource Misallocation

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

Table 1: Examples of duplicated AI safety efforts across organizations

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

1.2.3 Negative-Sum Dynamics

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

The game-theoretic structure of AI development creates particularly pernicious dynamics. Arm-

strong et al. **armstrong2016** demonstrate how uncoordinated policies can incentivize a "race to the precipice" where competitive pressures override safety considerations. The situation resembles a multi-player prisoner's dilemma or stag hunt where individually rational decisions lead to collectively catastrophic outcomes **samuel2023 hunt2025**.

1.3 Historical Parallels and Temporal Urgency

History offers instructive parallels. The nuclear age began with scientists racing to understand and control forces that could destroy civilization. Early coordination failures—competing national programs, scientist-military tensions, public-expert divides—nearly led to catastrophe multiple times. Only through developing shared frameworks (deterrence theory) **schelling1960**, institutions (IAEA), and communication channels (hotlines, treaties) did humanity navigate the nuclear precipice **rehman2025**.

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

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

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

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

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

Show Image

1.4 Research Question and Scope

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

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

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

To break this down into its components:

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

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

1.5 The Multiplicative Benefits Framework

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

Show Image

1.5.1 Automated Worldview Extraction

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

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

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

1.5.2 Live Data Integration

Static models, however well-constructed, quickly become outdated in fast-moving domains. Prediction markets and forecasting platforms aggregate distributed knowledge about uncertain futures, providing continuously updated probability estimates. By connecting formal models to these live data sources, we create dynamic assessments that incorporate the latest collective intelligence **tetlock2015**.

This integration serves multiple purposes:

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

1.5.3 Formal Policy Evaluation

Formal policy evaluation transforms static risk assessments into actionable guidance by modeling how specific interventions alter critical parameters. Using causal inference techniques **pearl2000 pearl2009**, we can assess not just the probability of adverse outcomes but how those probabilities change under different policy regimes.

This enables genuinely evidence-based policy development:

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

1.5.4 The Synergy

The multiplicative benefits emerge from the interactions between components:

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

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

1.6 Thesis Structure and Roadmap

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

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

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

Chapter 4: Discussion - Implications and Limitations critically examines technical limitations and failure modes, conceptual concerns about formalization, integration with existing governance frameworks, scaling challenges and opportunities, and broader implications for epistemic security.

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

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

Show Image

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

2. Context and Theoretical Foundations

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

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

2.1 AI Existential Risk: The Carlsmith Model

To ground our discussion in concrete terms, I examine Joseph Carlsmith's "Is Power-Seeking AI an Existential Risk?" as an exemplar of structured reasoning about AI catastrophic risk **carlsmith2022**. Carlsmith's analysis stands out for its explicit probabilistic decomposition of the path from current AI development to potential existential catastrophe.

2.1.1 Six-Premise Decomposition

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

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

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

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

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

mermaid

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

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

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

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

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

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

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

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

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

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

2.1.2 Why Carlsmith Exemplifies Formalizable Arguments

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

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

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

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

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

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

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

2.2 The Epistemic Challenge of Policy Evaluation

AI governance policy evaluation faces unique epistemic challenges that render traditional policy analysis methods insufficient. Understanding these challenges motivates the need for new computational approaches.

2.2.1 Unique Characteristics of AI Governance

Deep Uncertainty Rather Than Risk: Traditional policy analysis distinguishes between risk (known probability distributions) and uncertainty (known possibilities, unknown probabilities). AI governance faces deep uncertainty—we cannot confidently enumerate possible futures, much less assign probabilities **hallegatte2012**. Will recursive self-improvement enable rapid capability gains? Can value alignment be solved technically? These foundational questions resist empirical resolution before their answers become catastrophically relevant.

Complex Multi-Level Causation: Policy effects propagate through technical, institutional, and social levels with intricate feedback loops. A technical standard might alter research incentives, shifting capability development trajectories, changing competitive dynamics, and ultimately affecting existential risk through pathways invisible at the policy's inception. Traditional linear causal models cannot capture these dynamics.

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

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

Table 2: Comparison of AI governance vs traditional policy domains

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

2.2.2 Limitations of Traditional Approaches

Standard policy evaluation tools prove inadequate for these challenges:

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

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

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

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

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

2.2.3 The Underlying Epistemic Framework

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

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

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

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

This framework enables several key capabilities:

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

2.2.4 Toward New Epistemic Tools

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

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

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

Show Image

Show Image

Show Image

Show Image

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

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

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

2.3 Bayesian Networks as Knowledge Representation

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

2.3.1 Mathematical Foundations

A Bayesian network consists of:

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

The joint probability distribution factors according to the graph structure:

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

This factorization enables efficient inference and embodies causal assumptions explicitly.

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

2.3.2 The Rain-Sprinkler-Grass Example

The canonical example illustrates key concepts[3]:

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

Network Structure:

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

mermaid

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

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

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

python

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

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

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

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

This simple network demonstrates:

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

Show Image

Rain-Sprinkler-Grass Network Rendering

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

from IPython.display import IFrame

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

2.3.3 Advantages for AI Risk Modeling

These features address key requirements for AI governance:

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

Bibliography

UNIVERSITÄT BAYREUTH

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

Affidavit

Declaration of Academic Honesty

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

Automating the Modelling of Transformative Artificial Intelligence Risks

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

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

Bayreuth on the
May 26, 2025

Valentin Meyer