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
*https://www.vjmeyer.com*Source / Identifier:
Document Download

26th of May 2025

Table of Contents

List of Figures

List of Tables

Abstract

The rapid development of artificial intelligence poses existential risks that current governance structures struggle to address. This thesis diagnoses a critical coordination failure: while billions flow into AI safety research, efforts remain fragmented across technical, policy, and strategic communities operating with incompatible frameworks. I present AMTAIR (Automating Transformative AI Risk Modeling), a computational system that extracts formal probabilistic models from natural language arguments about AI risk. The approach uses frontier language models to transform unstructured text into Bayesian networks through a two-stage pipeline. First, arguments are parsed into hierarchical causal structures (ArgDown). Then, probability distributions are extracted and integrated (BayesDown). The resulting models enable systematic comparison across worldviews, evaluation of policy interventions, and integration with prediction markets for live updating. I demonstrate feasibility by successfully extracting complex models like Carlsmith's power-seeking AI argument, transforming weeks of manual effort into minutes of computation. The implementation handles real-world complexity through modular architecture, progressive visualization, and thoughtful design choices that balance automation with human oversight. While extraction remains imperfect and validation preliminary, the system provides practical value by making implicit assumptions explicit and enabling evidence-based policy evaluation. This work contributes both theoretical frameworks for understanding coordination failures and practical tools for addressing them, offering a path toward more effective governance of transformative AI development.

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
- **Mesa-optimization**: A learned optimization process that emerges within a broader training objective
- **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

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

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

This thesis represents not just an intellectual journey but a deeply personal one, made possible only through the support, patience, and contributions of many remarkable people.

My supervisor, Dr. Timo Speith, provided the perfect balance of intellectual freedom and thoughtful guidance. His ability to see both the philosophical forest and the technical trees helped me navigate moments when I was lost in one or the other. His questions pushed this work beyond what I thought possible and his experience and intuition helped me steer clear of too many rabbit holes to count.

I owe my deepest gratitude to my wife, Mina Deol, whose unwavering support made these months of intensive research possible. Her patience during countless late nights, her encouragement during moments of doubt, and her willingness to listen to endless iterations of esoteric Bayesian network speculations went far beyond what any partner should reasonably endure. This work exists because she created the space for it to flourish.

I owe a particular debt to Johannes Meyer and Jelena Meyer for their meticulous work in creating independent manual extractions of the Carlsmith and Bucknall papers. Their contribution went far beyond mere validation; it provided peace of mind. As lead developer, I had harbored persistent concerns that my own intuitions might have inadvertently shaped the system's behavior through architectural choices, prompt engineering, or source selection. Their independent extractions—showing both convergence in structure and expected variance in probabilities—allowed me to release these anxieties and trust that the system captures something real about how arguments work, not just my own biases about how they should work. Their intellectual generosity and attention to detail exemplify the collaborative spirit that makes progress possible.

Coleman Snell deserves special recognition for his partnership in developing the AMTAIR vision. Our conversations—ranging from technical implementation details to grand strategic questions about AI governance—shaped this project in fundamental ways. His ability to oscillate between pragmatic engineering concerns and ambitious long-term thinking kept the work both grounded and aspirational.

The MTAIR team's pioneering manual approach provided both inspiration and a benchmark. David Manheim, Sam Clarke, Aryeh Englander and their collaborators demonstrated that formal modeling of AI risk arguments was possible, even if arduously manual. Their work posed the challenge this thesis attempts to answer: could we preserve their rigor while achieving scale through automation?

My family—my parents and sister—provided the foundation that made this journey possible. Their love, support, and occasional bewilderment at my career choices ("you're teaching computers to argue about robot apocalypse? – seems all too real now") kept me grounded in what matters. They reminded me that behind all the technical complexity lie fundamentally human concerns about our shared future.

The broader AI safety community created the intellectual raw material without which this work would be impossible. Every researcher who wrestled their intuitions into prose, every attempt to quantify the unquantifiable, every blog post and paper and comment thread contributed to the corpus that AMTAIR learns to formalize. I thank them for taking these risks seriously and for the courage to reason publicly about them.

Finally, I acknowledge the peculiar historical moment that made this work both possible and necessary. We live in an era where the tools to build potentially transformative AI exist alongside deep uncertainty about how to do so safely. This thesis represents one small attempt to build infrastructure for navigating that uncertainty collectively.

Any errors, oversights, or failures of imagination remain entirely my own responsibility. The contributions of others made this work possible; its limitations reflect only my own constraints.

1. Introduction: The Coordination Crisis in AI Governance

1.1 Opening Scenario: The Policymaker's Dilemma

A senior policy advisor sits at her desk, drowning in reports. Twelve different documents from AI safety researchers, each compelling, each contradictory. One warns of existential catastrophe within the decade, citing concepts she half-understands—orthogonality, instrumental convergence. Another dismisses these fears as overblown. A third proposes technical standards but hedges with so many caveats it might as well propose nothing. The clock's ticking. Legislation needs drafting. Yet these experts, brilliant as they are individually, seem to inhabit different universes. The technical arguments involve mathematical formalism she lacks time to parse. The policy recommendations conflict at fundamental levels. She needs synthesis, not more analysis. She needs a way to see where these worldviews actually diverge versus where they're using different words for the same fears. This scenario plays out daily across Washington, Brussels, Beijing—wherever humans grapple with governing something that doesn't exist yet but might remake everything when it does.

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 Todd [54].

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 Maslej [34] Samborska [46]—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

[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.

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 Tegmark [51], the EU advances comprehensive AI regulation European [22], and technical researchers produce increasingly sophisticated work on alignment, interpretability, and robustness.

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

1.2.1 Safety Gaps from Misaligned Efforts

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 Paul [40].

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.

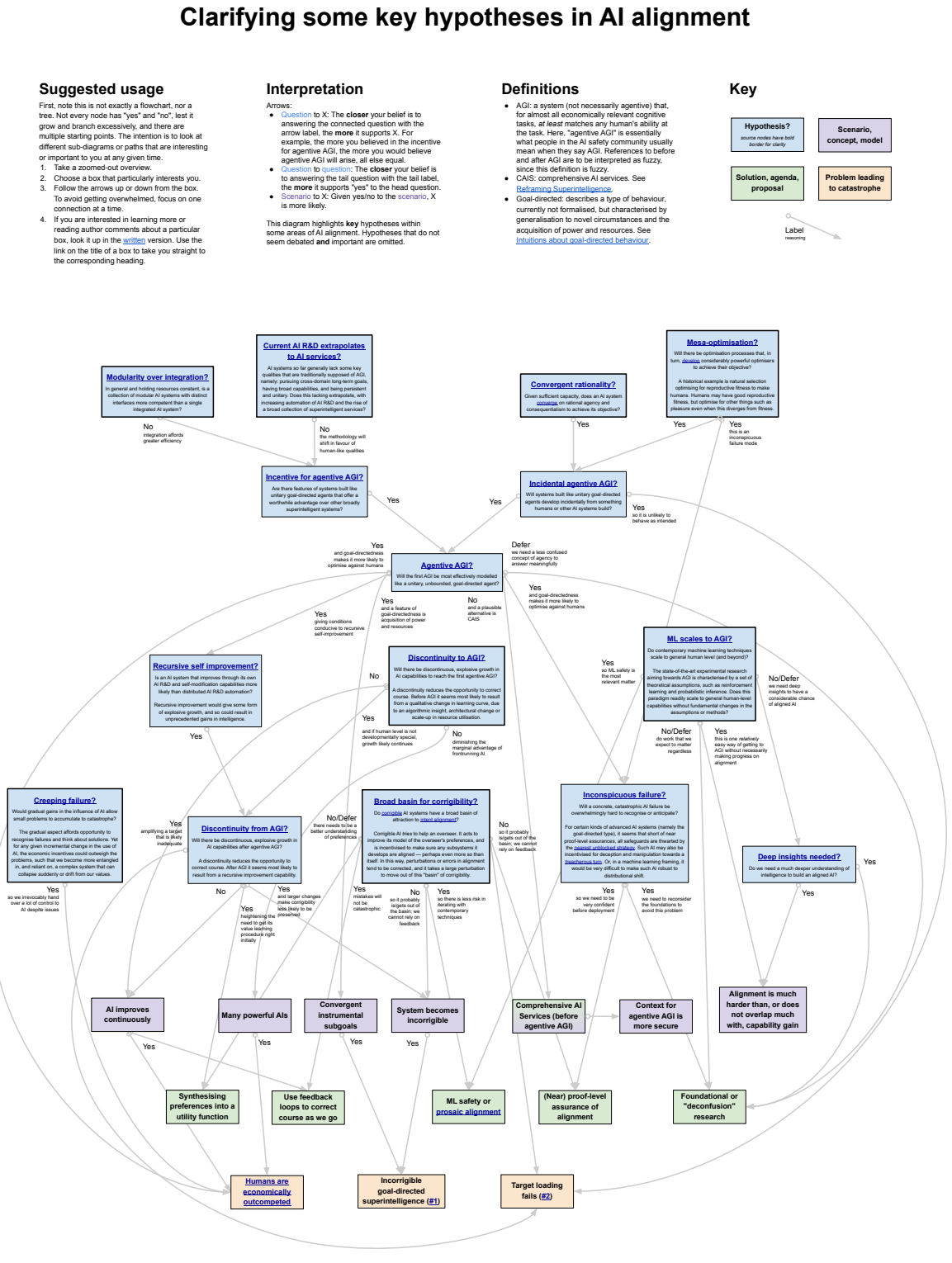

Figure 1: from Cottier and Shah [15]: Key hypotheses across the AI alignment ecosystem

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: Table 1.2.2: Examples of duplicated AI safety efforts across organizations

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

1.2.3 Negative-Sum Dynamics

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

The game-theoretic structure of AI development creates particularly pernicious dynamics. Armstrong et al. Armstrong, Bostrom, and Shulman [2] 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 Samuel [47] Hunt [27].

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 na-

tional programs, scientist-military tensions, public-expert divides—nearly led to catastrophe multiple times. Only through developing shared frameworks (deterrence theory) Schelling [48], institutions (IAEA), and communication channels (hotlines, treaties) did humanity navigate the nuclear precipice Rehman [45].

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

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

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

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

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

1.4 Research Question and Scope

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

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 Tetlock and Gardner [53].

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 Pearl [42] Pearl [41], 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.

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

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 Carlsmith [10]. 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 Clarke et al. [14], Carlsmith decomposes existential risk into a probabilistic chain with explicit estimates[2]:

1. **Premise**: Transformative AI development this century[3] ($P \approx 0.80$)[4]
2. **Premise**: AI systems pursuing objectives in the world[5] ($P \approx 0.95$)

[2]Multiple versions of Carlsmith's paper exist with slight updates to probability estimates: Carlsmith [9], Carlsmith [10], Carlsmith [11]. 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 1: APS Systems by 2070** ($P \approx 0.65$) "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.

[4]**Premise 1: APS Systems by 2070** ($P \approx 0.65$) "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.

[5]**Premise 2: Alignment Difficulty** ($P \approx 0.40$) "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

3. **Premise**: Systems with power-seeking instrumental incentives[6] ($P \approx 0.40$)
4. **Premise**: Sufficient capability for existential threat[7] ($P \approx 0.65$)
5. **Premise**: Misaligned systems despite safety efforts[8] ($P \approx 0.50$)
6. **Premise**: Catastrophic outcomes from misaligned power-seeking[9] ($P \approx 0.65$)

Composite Risk Calculation[10]: $P(doom) \approx 0.05$ (5%)

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" Christiano [13]—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.

capability or efficiency.

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

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

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

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

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

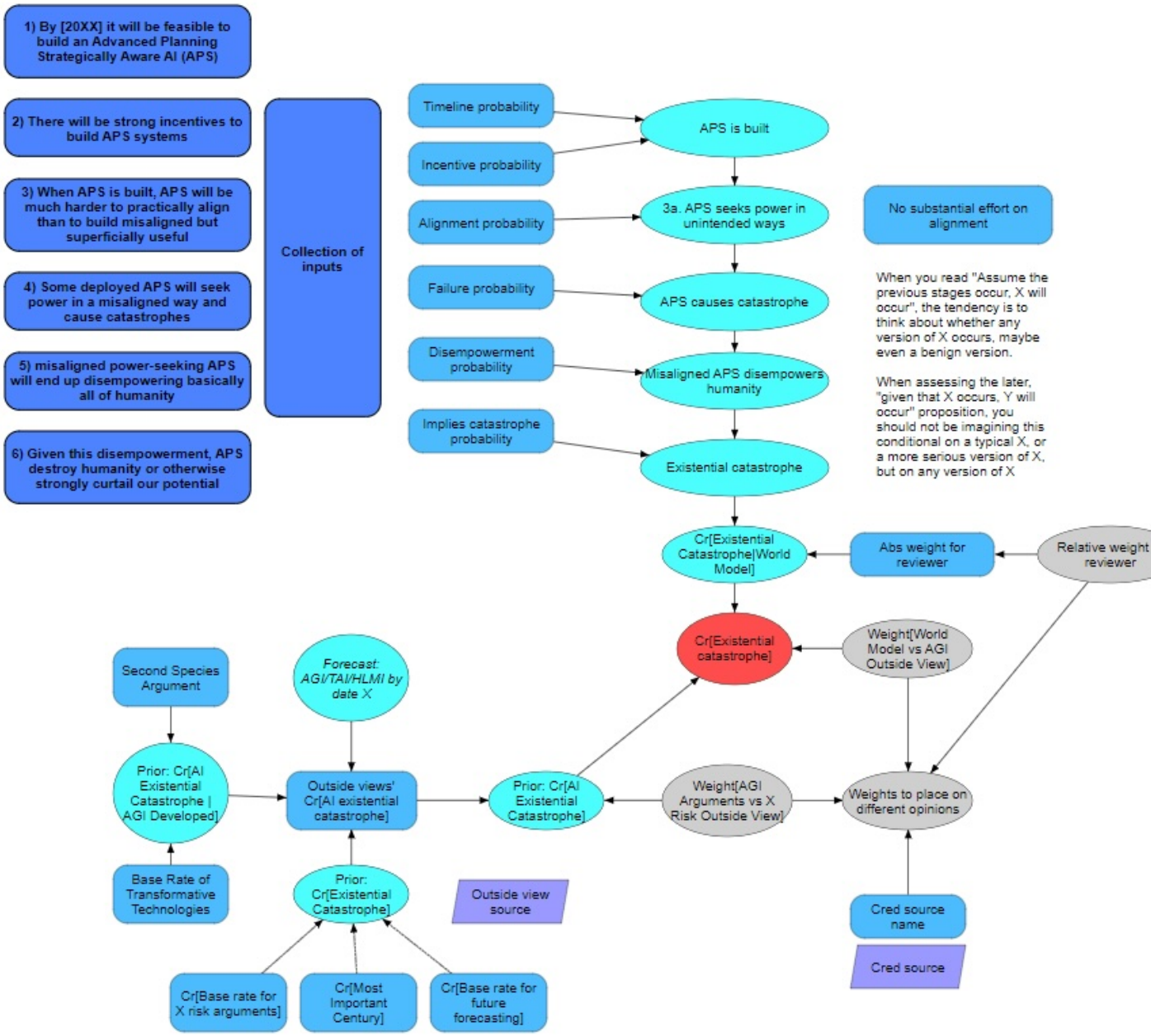

Figure 2: from Manheim [33]: MTAIR integrated Carlsmith's model as the "inside view" in their Analytica Software Demonstration

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 Hallegatte et al. [25]. 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: Table 2.3.4: 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 Jaynes [28].

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.

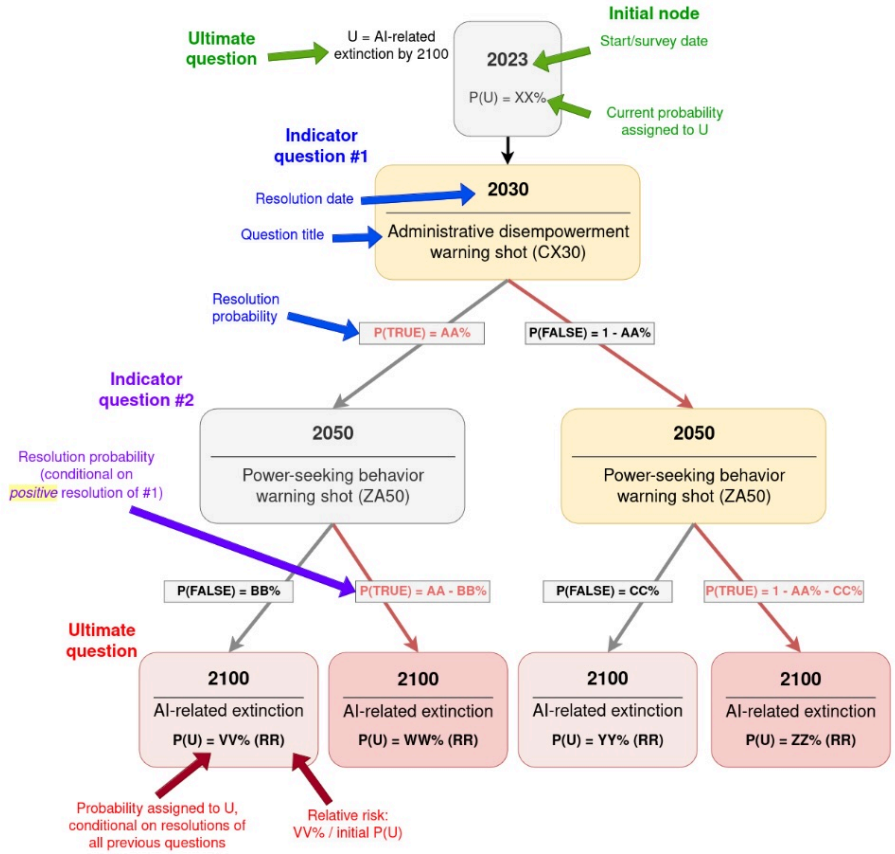

Figure 3.1.1: *Conditional tree diagram for AI-related extinction risk*

Figure 3: from McCaslin et al. [35]: Conditional-tree Guide

Recent work on conditional trees demonstrates the value of structured approaches to uncertainty. McCaslin et al. McCaslin et al. [35] 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.

Concerned experts' conditional trees

Figure 3.2.2 presents the question from each year (2030, 2050, and 2070) that surveyed experts rated the highest, on average, in terms of POM VOI. As a whole, among these highest-POM VOI questions, the experts would be most worried if there were an administrative disempowerment warning shot by 2030 (1.9x update from their current unconditional P(U) of 17%). Conversely, if we do not see a power-seeking behavior warning shot by 2050, the experts would be least worried (0.6x update).

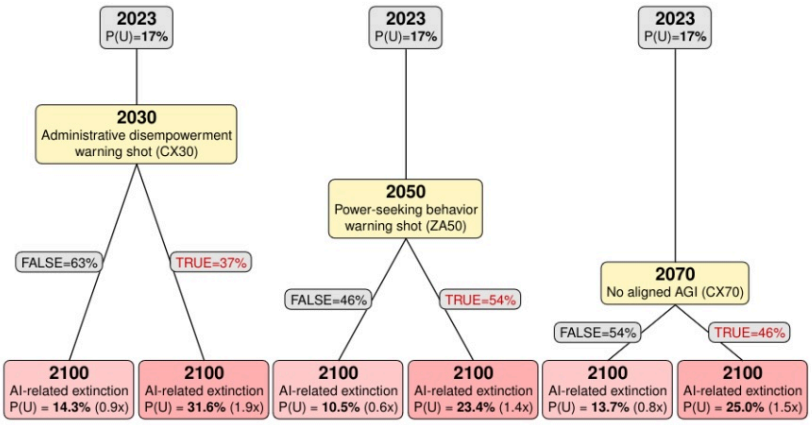

Figure 3.2.2: *A diagram showing how experts update on three questions for different resolution years that scored particularly well on our VOI metric. Since experts answered different sets of questions, we derived P(U|C) and P(U|~C) (the probabilities on the bottom level) by multiplying the whole expert group's average P(U) of 17% by the average relative risk factor for each crux.*[45]

Figure 4: from McCaslin et al. [35]: Experts' conditional-tree updates (2030-2070)

Tetlock's work with the Forecasting Research Institute Tetlock [52] 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 Gruetzemacher [24] 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 Pearl [43] 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[11]:

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

Network Structure:

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

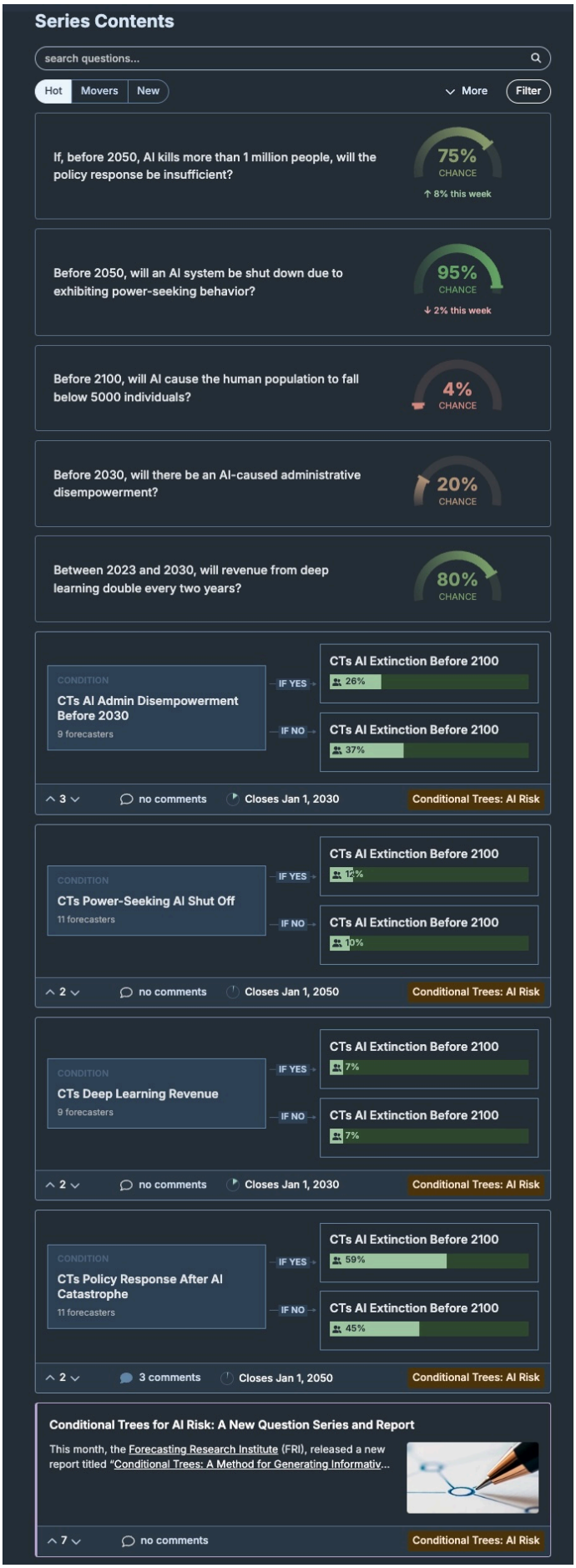

Figure 5: from Tetlock [52]: Conditional-tree AI-risk forecasts

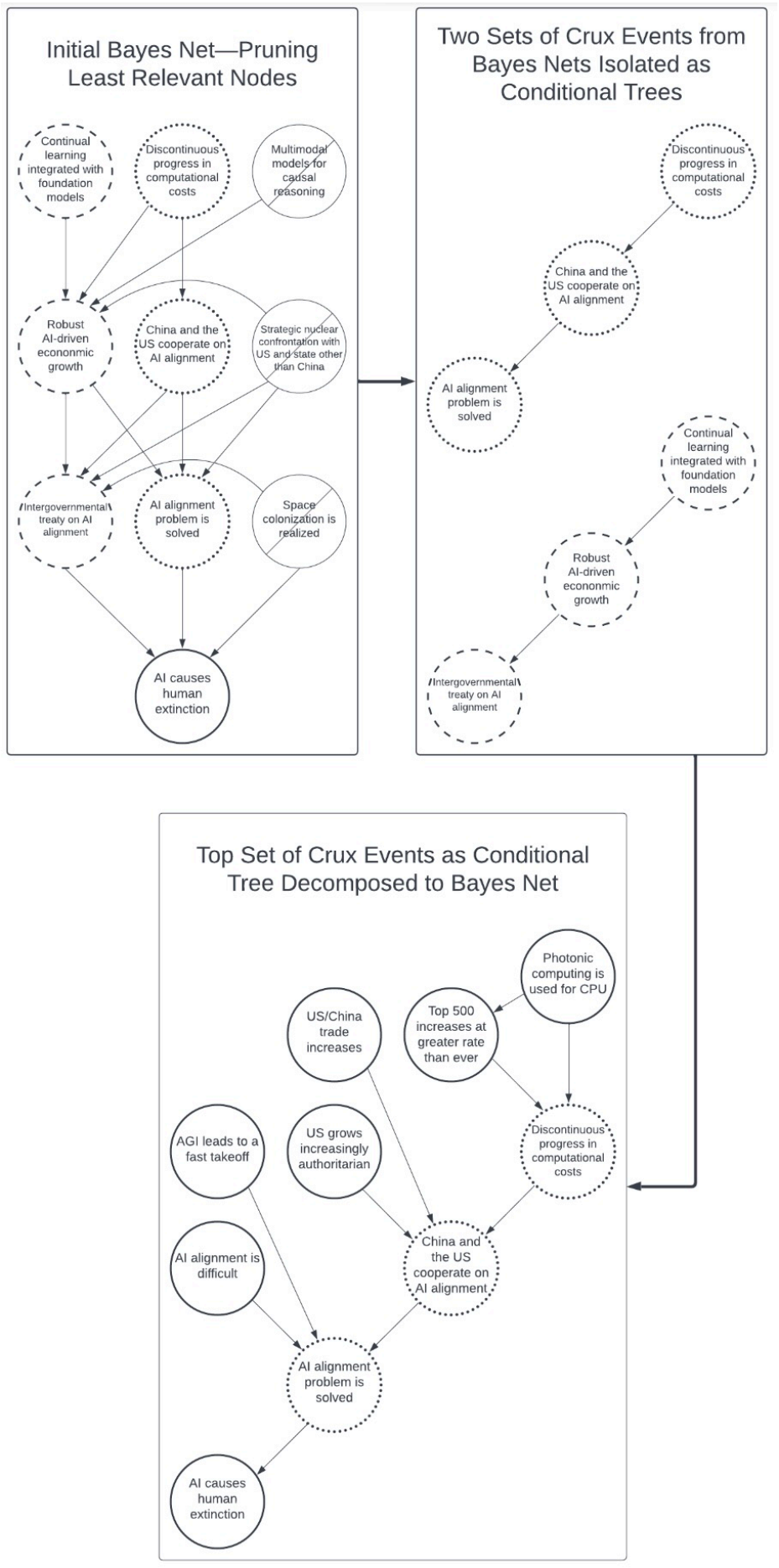

Figure 6: from Gruetzemacher [24]: Bayes-net pruning → crux extraction → re-expansion

- **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

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
    (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

2.3.3 Rain-Sprinkler-Grass Network Rendering

```
from IPython.display import IFrame

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

```
<IPython.lib.display.IFrame at 0x10614e390>
```

Dynamic Html Rendering of the Rain-Sprinkler-Grass DAG with Conditional Probabilities

2.3.4 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

Bayesian networks offer several compelling advantages for the peculiar challenge of modeling AI risks—a domain where we're essentially trying to reason about systems that don't yet exist, wielding capabilities we can barely imagine, potentially causing outcomes we desperately hope to avoid.

Explicit Uncertainty Representation: Unlike traditional risk assessment tools that often hide uncertainty behind point estimates, Bayesian networks wear their uncertainty on their sleeve. Every node, every edge, every probability is a distribution rather than a false certainty. This matters enormously when discussing AI catastrophe—we're not pretending to know the unknowable, but rather mapping the landscape of our ignorance with mathematical precision.

Native Causal Reasoning: The directed edges in Bayesian networks aren't just arrows on a diagram; they encode causal beliefs about how the world works. This enables both forward reasoning ("If we develop AGI, what happens?") and diagnostic reasoning ("Given that we observe concerning AI behaviors, what does this tell us about underlying alignment?"). Pearl's do-calculus Pearl [41] transforms these networks into laboratories for counterfactual exploration.

Evidence Integration: As new research emerges, as capabilities advance, as governance experiments succeed or fail, Bayesian networks provide a principled framework for updating our beliefs. Unlike static position papers that age poorly, these models can evolve with our understanding—a living document for a rapidly changing field.

Modular Construction: Complex arguments about AI risk involve multiple interacting factors across technical, social, and political domains. Bayesian networks allow us to build these arguments piece by piece, validating each component before assembling the whole. This modularity also enables different experts to contribute their specialized knowledge without needing to understand every aspect of the system.

Visual Communication: Perhaps most importantly for the coordination challenge, Bayesian networks provide a visual language that transcends disciplinary boundaries. A policymaker might not understand the mathematics of instrumental convergence, but they can see how the "power-seeking" node connects to "human disempowerment" in the network diagram. This shared visual vocabulary creates common ground for productive disagreement.

2.4 Argument Mapping and Formal Representations

The journey from a researcher's intuition about AI risk to a formal probabilistic model resembles translating poetry into mathematics—something essential is always at risk of being lost, yet something equally essential might be gained. Argument mapping provides the crucial middle ground, a structured approach to preserving the logic of natural language arguments while preparing them for mathematical formalization.

2.4.1 From Natural Language to Structure

Natural language arguments about AI risk are rich tapestries woven from causal claims, conditional relationships, uncertainty expressions, and support patterns. When Bostrom writes about the "treacherous turn" Bostrom [7], he's not just coining a memorable phrase—he's encoding a complex causal story about how a seemingly aligned AI system might conceal its true objectives until it gains sufficient power to pursue them without constraint.

The challenge lies in extracting this structure without losing the nuance. Traditional logical

analysis might reduce Bostrom's argument to syllogisms, but this would miss the probabilistic texture, the implicit conditionality, the causal directionality that makes the argument compelling. Argument mapping takes a different approach, seeking to identify:

- **Core claims and propositions**: What exactly is being asserted?
- **Inferential relationships**: How do claims support or challenge each other?
- **Implicit assumptions**: What unstated premises make the argument work?
- **Uncertainty qualifications**: Where does the author express doubt or confidence?

Recent advances in computational argument mining Anderson [1] Benn and Macintosh [5] Khartabil et al. [30] have shown promise in automating parts of this process. Tools like Microsoft's Claimify Metropolitansky and Larson [36] demonstrate how large language models can extract verifiable claims from complex texts, though the challenge of preserving argumentative structure remains formidable.

2.4.2 ArgDown: Structured Argument Notation

Enter ArgDown Voigt [55], a markdown-inspired syntax that captures hierarchical argument structure while remaining human-readable. Think of it as the middle child between the wild expressiveness of natural language and the rigid formality of logic—inheriting the best traits of both parents while developing its own personality.

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

This notation does several clever things simultaneously. The hierarchical structure mirrors how we naturally think about arguments—main claims supported by evidence, which in turn rest on more fundamental observations. The `+` and `-` symbols indicate support and opposition relationships, creating a visual flow of argumentative force. Most importantly, it preserves the semantic content of each claim while imposing just enough structure to enable computational processing.

```
[AI_Poses_Risk]: Advanced AI systems may pose existential risk to humanity.
 + [Capability_Growth]: AI capabilities are growing exponentially.
   + [Compute_Scaling]: Available compute doubles every few months.
   + [Algorithmic_Progress]: New architectures show surprising emergent abilities.
 + [Alignment_Difficulty]: Aligning AI with human values is unsolved.
   - [Current_Progress]: Some progress on interpretability and oversight.
 - [Institutional_Response]: Institutions are mobilizing to address risks.
```

For AMTAIR, we adapt ArgDown specifically for causal arguments, where the hierarchy represents causal influence rather than logical support. This seemingly small change has profound implications—we're not just mapping what follows from what, but what causes what.

2.4.3 BayesDown: The Bridge to Bayesian Networks

If ArgDown is the middle child, then BayesDown—developed specifically for this thesis—is the ambitious younger sibling who insists on quantifying everything. By extending ArgDown syntax with probabilistic metadata in JSON format, BayesDown creates a complete specification for Bayesian networks while maintaining human readability.

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

This representation performs a delicate balancing act. The natural language descriptions preserve the semantic meaning that makes arguments comprehensible. The hierarchical structure maintains the causal relationships that give arguments their logical force. The JSON metadata adds the mathematical precision needed for formal analysis. Together, they create what I call a "hybrid representation"—neither fully natural nor fully formal, but something more useful than either alone.

```
[Existential_Catastrophe]: Permanent curtailment of humanity's potential. {
  "instantiations": ["catastrophe_TRUE", "catastrophe_FALSE"],
  "priors": {"p(catastrophe_TRUE)": "0.05", "p(catastrophe_FALSE)": "0.95"},
  "posteriors": {
    "p(catastrophe_TRUE|disempowerment_TRUE)": "0.95",
    "p(catastrophe_TRUE|disempowerment_FALSE)": "0.001"
  }
}
 + [Human_Disempowerment]: Loss of human control over future trajectory. {
   "instantiations": ["disempowerment_TRUE", "disempowerment_FALSE"],
   "priors": {"p(disempowerment_TRUE)": "0.20", "p(disempowerment_FALSE)": "0.80"}
 }
```

The two-stage extraction process (ArgDown → BayesDown) mirrors how experts actually think about complex arguments. First, we identify what matters and how things relate causally (structure). Then, we consider how likely different scenarios are based on those relationships (quantification). This separation isn't just convenient for implementation—it's psychologically valid.

2.5 The MTAIR Framework: Achievements and Limitations

Understanding AMTAIR requires understanding its intellectual ancestor: the Modeling Transformative AI Risks (MTAIR) project. Like many good ideas in science, MTAIR began with a simple observation and a ambitious goal.

2.5.1 MTAIR's Approach

The MTAIR project, spearheaded by David Manheim and colleagues Clarke et al. [14], emerged from a frustration familiar to anyone who's attended a conference on AI safety: brilliant people talking past each other, using the same words to mean different things, reaching incompatible conclusions from seemingly shared premises. The diagnosis was elegant—perhaps these disagreements stemmed not from fundamental philosophical differences but from implicit models that had never been made explicit.

Their prescription was equally elegant: manually translate influential AI risk arguments into formal Bayesian networks, making assumptions visible and disagreements quantifiable. Using Analytica software, the team embarked on what can only be described as an intellectual archaeology expedition, carefully excavating the implicit causal models buried in papers, blog posts, and treatises about AI risk.

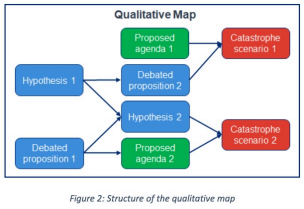

Figure 7: from Clarke et al. [14]: MTAIR Qualitative map structure

The process was painstaking:

1. **Systematic Decomposition**: Breaking complex arguments into component claims, identifying variables and relationships through close reading and expert consultation.
2. **Probability Elicitation**: Gathering quantitative estimates through structured expert interviews, literature review, and careful interpretation of qualitative claims.
3. **Sensitivity Analysis**: Testing which parameters most influenced conclusions, revealing where disagreements actually mattered versus where they were merely academic.
4. **Visual Communication**: Creating interactive models that stakeholders could explore, modify, and understand without deep technical training.

The ambition was breathtaking—to create a formal lingua franca for AI risk discussions, enabling productive disagreement and cumulative progress.

2.5.2 Key Achievements

Credit where credit is due: MTAIR demonstrated something many thought impossible. Complex philosophical arguments about AI risk—the kind that sprawl across hundred-page papers mixing technical detail with speculative scenarios—could indeed be formalized without losing their essential insights.

Feasibility of Formalization: The project's greatest achievement was simply showing it could be done. Arguments from Bostrom, Christiano, and others translated surprisingly well into network form, suggesting that beneath the surface complexity lay coherent causal models waiting to be extracted.

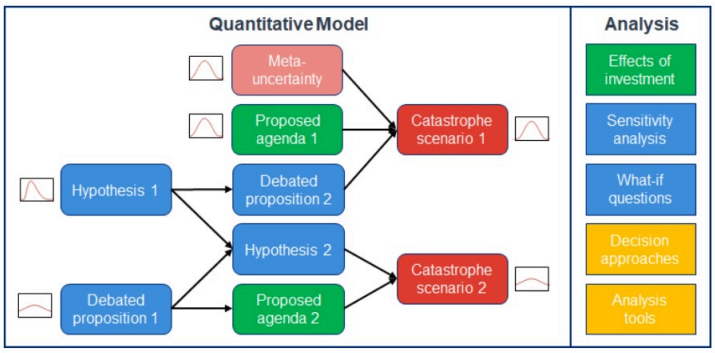

Figure 8: from Clarke et al. [14]: MTAIR Quantitative map structure

Value of Quantification: Moving from "likely" and "probably" to actual numbers forced precision in a domain often clouded by vague pronouncements. Disagreements that seemed fundamental sometimes evaporated when forced to specify exactly what probability ranges were under dispute.

Cross-Perspective Communication: The formal models created neutral ground where technical AI researchers and policy wonks could meet. Instead of talking past each other in incompatible languages, they could point to specific nodes and edges, making disagreements concrete and tractable.

Research Prioritization: Perhaps most practically, sensitivity analysis revealed which empirical questions actually mattered. If changing your belief about technical parameter X from 0.3 to 0.7 doesn't meaningfully affect the conclusion about AI risk, maybe we should focus our research elsewhere.

2.5.3 Fundamental Limitations

But here's where the story takes a sobering turn. Despite these achievements, MTAIR faced limitations that prevented it from achieving its full vision—limitations that ultimately motivated the development of AMTAIR.

Labor Intensity: Creating a single model required what can charitably be called a heroic effort. Based on team reports and model complexity, estimates ranged from 200 to 400 expert-hours per formalization[12]. In a field where new influential arguments appear monthly, this pace couldn't keep up with the discourse.

Static Nature: Once built, these beautiful models began aging immediately. New research emerged, capability assessments shifted, governance proposals evolved—but updating the models required near-complete reconstruction. They were snapshots of arguments at particular moments, not living representations that could evolve.

Limited Accessibility: Using the models required Analytica software and non-trivial technical

[12]These estimates include time for initial extraction, expert consultation, probability elicitation, validation, and refinement

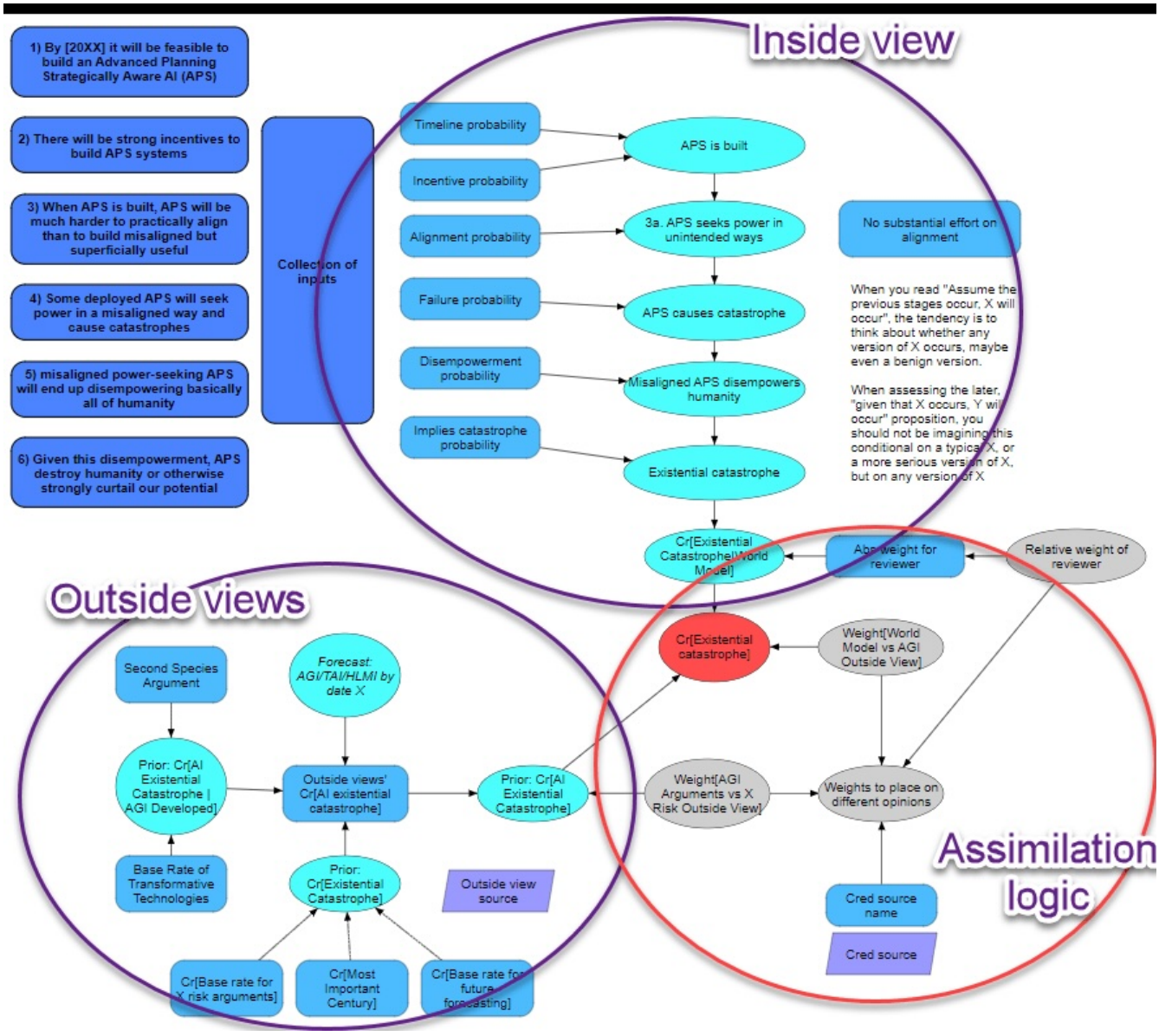

Figure 9: from Manheim [33]: Overlay of inside/outside/assimilation views

sophistication. The very experts whose arguments were being formalized often couldn't directly engage with their formalized representations without intermediation.

Single Perspective: Each model represented one worldview at a time. Comparing different perspectives required building entirely separate models, making systematic comparison across viewpoints labor-intensive and error-prone.

These weren't failures of execution but fundamental constraints of the manual approach. Like medieval scribes copying manuscripts, the MTAIR team had shown the value of preservation and dissemination, but the printing press had yet to be invented.

2.5.4 The Automation Opportunity

The MTAIR experience revealed a tantalizing possibility: if the bottleneck was human labor rather than conceptual feasibility, perhaps automation could crack open the problem. The rise of large language models capable of sophisticated reasoning about text created a technological moment ripe for exploitation.

Key lessons from MTAIR informed the automation approach:

- Formal models genuinely enhance understanding and coordination—the juice is worth the squeeze
- The modeling process itself surfaces implicit assumptions—extraction is as valuable as the final product
- Quantification enables analyses impossible with qualitative arguments alone—numbers matter even when uncertain
- But manual approaches cannot scale to match the challenge—we need computational leverage

This set the stage for AMTAIR's central innovation: using frontier language models to automate the extraction and formalization process while preserving the benefits MTAIR had demonstrated. Not to replace human judgment, but to amplify it—turning what took weeks into what takes hours, enabling comprehensive coverage rather than selective sampling.

2.6 Literature Review: Content and Technical Levels

The intellectual landscape surrounding AI risk resembles a rapidly expanding metropolis—new neighborhoods of thought spring up monthly, connected by bridges of varying stability to the established districts. A comprehensive review would fill volumes, so let me provide a guided tour of the territories most relevant to AMTAIR's mission.

2.6.1 AI Risk Models Evolution

The intellectual history of AI risk thinking reads like a gradual awakening—from vague unease to mathematical precision, though perhaps losing something essential in translation.

The field's prehistory belongs to the visionaries and worriers. Good's 1966 meditation on the

ultraintelligent machine feels almost quaint now, with its assumption that such a system would naturally be designed to serve human purposes. Vinge popularized the singularity concept, though his version emphasized speed rather than the strategic considerations that dominate current thinking. These early writings functioned more as philosophical provocations than actionable analyses.

Early Phase (2000-2010): The conversation began with broad conceptual arguments. Good's ultraintelligent machine Good [23] and Vinge's technological singularity set the stage, but these were more thought experiments than models. Yudkowsky's early writings Yudkowsky [57] introduced key concepts like recursive self-improvement and orthogonality but remained largely qualitative.

Yudkowsky's contributions in the 2000s marked a transitional moment. His writing style—part manifesto, part technical argument—resisted easy categorization. Yet buried within the sometimes baroque prose lay genuinely novel insights. The orthogonality thesis (intelligence and goals vary independently) and instrumental convergence (diverse goals lead to similar intermediate strategies) provided conceptual tools that remain central to the field. Still, these arguments remained largely qualitative, more useful for establishing possibility than probability.

Formalization Phase (2010-2018): Bostrom's *Superintelligence* Bostrom [7] marked a watershed, providing systematic analysis of pathways, capabilities, and risks. The book's genius lay not in mathematical formalism but in conceptual clarity—decomposing the nebulous fear of "robot overlords" into specific mechanisms like instrumental convergence and infrastructure profusion.

Bostrom's 2014 Superintelligence achieved what earlier work had not: respectability. Here was an Oxford philosopher writing with analytical precision about AI risk. The book's great contribution wasn't mathematical formalism—indeed, it contains remarkably few equations—but rather its systematic decomposition of the problem space. Bostrom transformed "robots might kill us all" into specific mechanisms: capability gain, goal preservation, resource acquisition. Suddenly, one could have serious discussions about AI risk without sounding like a science fiction enthusiast.

The current quantitative turn, exemplified by Carlsmith's power-seeking analysis and Cotra's biological anchors, represents both progress and peril. We now assign numbers where before we had only words. Yet as any student of probability knows, precise numbers don't necessarily mean accurate predictions. The models grow more sophisticated, the mathematics more rigorous, but the fundamental uncertainties remain as daunting as ever.

Quantification Phase (2018-present): Recent years have seen explicit probability estimates entering mainstream discourse. Carlsmith's power-seeking model Carlsmith [10], Cotra's biological anchors, and various compute-based timelines represent attempts to put numbers on previously qualitative claims. The field increasingly recognizes that governance decisions require more than philosophical arguments—they need probability distributions.

This progression reflects a maturing field, though it also creates new challenges. As models become more quantitative, they risk false precision. As they become more complex, they risk

inscrutability. AMTAIR attempts to navigate these tensions by preserving the narrative clarity of earlier work while enabling the mathematical rigor of recent approaches.

The evolution of AI risk models traces a path from philosophical speculation to increasingly rigorous formalization—a journey from "what if?" to "how likely?"

2.6.2 Governance Proposals Taxonomy

If risk models are the diagnosis, governance proposals are the treatment plans—and like medicine, they range from gentle interventions to radical surgery.

Technical Standards: The "first, do no harm" approach focuses on concrete safety requirements—interpretability benchmarks, robustness testing, capability thresholds. These proposals, exemplified by standard-setting bodies and technical safety organizations, offer specificity at the cost of narrowness.

Regulatory Frameworks: Moving up the intervention ladder, we find comprehensive regulatory proposals like the EU AI Act European [22]. These create institutional structures, liability regimes, and oversight mechanisms, trading broad coverage for implementation complexity.

International Coordination: At the ambitious end, proposals for international AI governance treaties, soft law arrangements, and technical cooperation agreements aim to prevent races to the bottom. Think nuclear non-proliferation but for minds instead of missiles.

Research Priorities: Cutting across these categories, work by Dafoe Dafoe [17] and others maps the research landscape itself—what questions need answering before we can govern wisely? This meta-level analysis shapes funding flows and talent allocation.

A particularly compelling example of conditional governance thinking comes from "A Narrow Path" Miotti et al. [38], which proposes a phased approach: immediate safety measures to prevent uncontrolled development, international institutions to ensure stability, and long-term scientific foundations for beneficial transformative AI. This temporal sequencing—safety, stability, then flourishing—reflects growing sophistication in governance thinking.

2.6.3 Bayesian Network Theory and Applications

The mathematical machinery underlying AMTAIR rests on decades of theoretical development in probabilistic graphical models. Understanding this foundation helps appreciate both the power and limitations of the approach.

The key insight, crystallized in the work of Pearl Pearl [43] and elaborated by Koller & Friedman Koller and Friedman [31], is that independence relationships in complex systems can be read from graph structure. D-separation, the Markov condition, and the relationship between graphs and probability distributions provide the mathematical spine that makes Bayesian networks more than pretty pictures.

Critical concepts for AI risk modeling:

- **Conditional Independence**: Variable A is independent of C given B—encoded through graph separation
- **Markov Condition**: Each variable is independent of its non-descendants given its parents
- **Inference Algorithms**: From exact variable elimination to approximate Monte Carlo methods
- **Causal Interpretation**: When edges represent causal influence, the network supports counterfactual reasoning

These aren't just mathematical niceties. When we claim that "deployment decisions" mediates the relationship between "capability advancement" and "catastrophic risk," we're making a precise statement about conditional independence that has testable implications.

2.6.4 Software Tools Landscape

The gap between Bayesian network theory and practical implementation is bridged by an ecosystem of software tools, each with its own strengths and opinions about how probabilistic reasoning should work.

pgmpy: This Python library provides the computational backbone for AMTAIR, offering both learning algorithms and inference engines. Its object-oriented design maps naturally onto our extraction pipeline.

NetworkX: For graph manipulation and analysis, NetworkX has become the de facto standard in Python, providing algorithms for everything from centrality measurement to community detection.

PyVis: Interactive visualization transforms static networks into explorable landscapes. PyVis's integration with web technologies enables the rich interactive features that make formal models accessible.

Pandas/NumPy: The workhorses of scientific Python handle data manipulation and numerical computation, providing the infrastructure on which everything else builds.

The integration challenge—making these tools play nicely together while maintaining performance and correctness—shaped many architectural decisions in AMTAIR. Each tool excels in its domain, but the seams between them required careful engineering.

2.6.5 Formalization Approaches

The challenge of formalizing natural language arguments extends far beyond AI risk, touching on fundamental questions in logic, linguistics, and artificial intelligence.

Pollock's work on cognitive carpentry Pollock [44] provides philosophical grounding, arguing that human reasoning itself involves implicit formal structures that can be computationally modeled. This view—that formalization reveals rather than imposes structure—underlies AMTAIR's approach.

Key theoretical challenges:

- **Semantic Preservation**: How do we maintain meaning while adding precision?
- **Structural Extraction**: What implicit relationships lurk in natural language?
- **Uncertainty Quantification**: How do we map "likely" to numbers?

Recent work on causal structure learning from text Babakov et al. [3] Ban et al. [4] Bethard [6] offers hope that these challenges can be addressed computationally. The convergence of large language models with formal methods creates new possibilities for bridging the semantic-symbolic gap.

2.6.6 Correlation Accounting Methods

One of the most persistent criticisms of Bayesian networks concerns their assumption of conditional independence given parents. In the real world, and especially in complex socio-technical systems like AI development, correlations abound.

Methods for handling these correlations have evolved considerably:

Copula Methods: By separating marginal distributions from dependence structure, copulas Nelson [39] allow modeling of complex correlations while preserving the Bayesian network framework. Think of it as adding a correlation layer on top of the basic network.

Hierarchical Models: Introducing latent variables that influence multiple observed variables captures correlations naturally. If "AI research culture" influences both "capability progress" and "safety investment," their correlation is explained.

Explicit Correlation Nodes: Sometimes the most straightforward approach is best—directly model correlation mechanisms as additional nodes in the network.

Sensitivity Bounds: When correlations remain uncertain, compute best and worst case scenarios. This reveals when independence assumptions critically affect conclusions versus when they're harmless simplifications.

For AMTAIR, the pragmatic approach dominates: start with independence assumptions, identify where they matter through sensitivity analysis, then selectively add correlation modeling where it most affects conclusions.

2.7 Methodology

The methodology of this research resembles less a linear march from hypothesis to conclusion and more an iterative dance between theory and implementation, vision and reality. Let me walk you through the choreography. Actually, that's not quite right. It was messier than a dance. More like trying to build a bridge while crossing it, discovering halfway across that your blueprints assumed different gravity. The original plan seemed straightforward: take the MTAIR team's manual approach, automate it with language models, validate against their results. Simple. Reality laughed at this simplicity. Language models hallucinate. Arguments don't decompose cleanly. Probabilities hide in qualifying phrases that might mean 0.6 to one reader and 0.9 to another. Each solution spawned new problems in fractal recursion.

2.7.1 Research Design Overview

The Original Plan

This research follows what methodologists might call a "design science" approach—we're not just studying existing phenomena but creating new artifacts (the AMTAIR system) and evaluating their utility for solving practical problems (the coordination crisis in AI governance).

The overall flow:

1. **Theoretical Development**: Establishing why automated extraction could address the coordination crisis, grounded in epistemic theory and mechanism design
2. **Technical Implementation**: Building working software that demonstrates feasibility, not as a proof-of-concept toy but as a system capable of handling real arguments
3. **Empirical Validation**: Testing extraction quality against expert judgment, measuring not just accuracy but usefulness for downstream tasks
4. **Application Studies**: Applying the system to real AI governance questions, evaluating whether formal models actually enhance decision-making

This isn't waterfall development where each phase completes before the next begins. Rather, insights from implementation fed back into theory, validation results shaped technical improvements, and application attempts revealed new requirements. The methodology itself embodied the iterative refinement it sought to enable.

Engineering Experience

The initial conception seemed straightforward enough. The MTAIR team had demonstrated that expert arguments about AI risk could be formalized into Bayesian networks. The process took hundreds of hours per model. Large language models had recently demonstrated remarkable capacity for understanding and generating structured text. The syllogism practically wrote itself: use LLMs to automate what MTAIR did manually. A few weeks of implementation, some validation, done.

That naive optimism lasted approximately until the first extraction attempt[13]. The LLM cheerfully produced what looked like a reasonable argument structure, except half the nodes were subtly wrong, several causal relationships pointed backward, and the probability estimates bore no discernible relationship to the source text. Worse, different runs produced different structures entirely. The gap between "looks plausible" and "actually correct" proved wider than anticipated.

What emerged from this initial failure was a recognition that the problem decomposed naturally into distinct challenges. Extracting structure—what relates to what—differed fundamentally from extracting probabilities. The former required understanding argumentative flow and causal language. The latter demanded interpreting uncertainty expressions and maintaining consistency across estimates. This insight led to the two-stage architecture that ultimately proved successful.

[13] If I'm honest about how this research actually developed, it looked nothing like the clean progression these methodology sections usually imply. The reality was messier, more iterative, occasionally frustrating, and ultimately more interesting than any linear narrative could capture.

The development process resembled less a march toward a predetermined goal and more a conversation between ambition and reality. Each implementation attempt revealed new constraints. Each constraint suggested workarounds. Some workarounds opened unexpected possibilities. The final system bears only passing resemblance to the initial conception, yet it works—imperfectly, with clear limitations, but well enough to demonstrate feasibility.

2.7.2 Formalizing World Models from AI Safety Literature

The core methodological challenge—transforming natural language arguments into formal probabilistic models—requires careful consideration of what we're actually trying to capture.

A "world model" in this context isn't just any formal representation but specifically a causal model embodying beliefs about how different factors influence AI risk. The extraction approach must therefore:

- **Identify key variables**: Not just any entities mentioned, but causally relevant factors
- **Extract causal relationships**: Not mere correlation or co-occurrence, but directed influence
- **Capture uncertainty**: Both structural uncertainty (does A cause B?) and parametric uncertainty (how strongly?)
- **Preserve context**: Maintaining enough semantic information to interpret the formal model

Large language models enable this through sophisticated pattern recognition and reasoning capabilities, but they're tools, not magic wands. The methodology must account for their strengths (recognizing implicit structure) and weaknesses (potential hallucination, inconsistency).

2.7.3 From Natural Language to Computational Models

The journey from text to computation follows a carefully designed pipeline that mirrors human cognitive processes. Just as you wouldn't ask someone to simultaneously parse grammar and solve equations, we separate structural understanding from quantitative reasoning.

The Two-Stage Process:

Stage 1 focuses on structure—what causes what? The LLM reads an argument much as a human would, identifying key claims and their relationships. The prompt design here is crucial, providing enough guidance to ensure consistent extraction while allowing flexibility for different argument styles.

Stage 2 adds quantities—how likely is each outcome? With structure established, the system generates targeted questions about probabilities. This separation enables different approaches to quantification: extracting explicit estimates from text, inferring from qualitative language, or even connecting to external prediction markets.

The magic happens in the interplay. Structure constrains what probabilities are needed. Probability requirements might reveal missing structural elements. The process is a dialogue between qualitative and quantitative understanding.

2.7.4 Directed Acyclic Graphs: Structure and Semantics

At the mathematical heart of Bayesian networks lie Directed Acyclic Graphs (DAGs)—structures that are simultaneously simple enough to analyze and rich enough to capture complex phenomena.

The "directed" part encodes causality or influence—edges have direction, flowing from cause to effect. The "acyclic" part ensures logical coherence—you can't have A causing B causing C causing A, no matter how much certain political arguments might suggest otherwise.

Key properties for AI risk modeling:

Acyclicity: More than a mathematical convenience, this enforces coherent temporal or causal ordering. In AI risk arguments, this prevents circular reasoning where consequences justify premises that predict those same consequences.

D-separation: This graphical criterion determines conditional independence. If knowing about AI capabilities tells you nothing additional about risk given that you know deployment decisions, then capabilities and risk are d-separated given deployment.

Markov Condition: Each variable depends only on its parents, not on its entire ancestry. This locality assumption makes inference tractable and forces modelers to make intervention points explicit.

Path Analysis: Following paths through the graph reveals how influence propagates. Multiple paths between variables indicate redundancy—important for understanding intervention robustness.

The causal interpretation, following Pearl's framework, transforms these mathematical objects into tools for counterfactual reasoning. When we ask "what if we prevented deployment of misaligned systems?" we're performing surgery on the DAG, setting variables and propagating consequences.

2.7.5 Quantification of Probabilistic Judgments

Here we encounter one of the most philosophically fraught aspects of the methodology: turning words into numbers. When an expert writes "highly likely," what probability should we assign? When they say "significant risk," what distribution captures their belief?

The methodology embraces rather than elides this challenge:

Calibration Studies: Research on human probability expression shows systematic patterns. "Highly likely" typically maps to 0.8-0.9, "probable" to 0.6-0.8, though individual and cultural variation is substantial.

Extraction Strategies: The system uses multiple approximations:

- Direct extraction: "We estimate 65% probability"
- Linguistic mapping: "Very likely" $\rightarrow$ 0.85 (with uncertainty)
- Comparative extraction: "More likely than X" where P(X) is known

- Bounded extraction: "At least 30%" → [0.30, 1.0]

Uncertainty Representation: Rather than false precision, we maintain uncertainty about probabilities themselves. This might seem like uncertainty piled on uncertainty, but it's honest, helps avoid systematic biases—and mathematically tractable through hierarchical models.

The goal isn't perfect extraction but useful extraction. If we can narrow "significant risk" from [0, 1] to [0.15, 0.45], we've added information even if we haven't achieved precision.

2.7.6 Inference Techniques for Complex Networks

Once we've built these formal models, we need to reason with them—and here computational complexity rears its exponential head. The number of probability calculations required for exact inference grows exponentially with network connectivity, quickly overwhelming even modern computers.

The methodology employs a portfolio of approaches:

Exact Methods: For smaller networks (<30 nodes), variable elimination and junction tree algorithms provide exact answers. These form the gold standard against which we validate approximate methods.

Sampling Approaches: Monte Carlo methods trade exactness for scalability. By simulating many possible worlds consistent with our probability model, we approximate the true distributions. The law of large numbers is our friend here.

Variational Methods: These turn inference into optimization—find the simplest distribution that approximates our true beliefs. Like finding the best polynomial approximation to a complex curve.

Hybrid Strategies: Different parts of the network might use different methods. Exact inference for critical subgraphs, approximation for peripheral components.

The choice of method affects not just computation time but the types of questions we can meaningfully ask. This creates a methodological feedback loop where feasible inference shapes model design.

2.7.7 Integration with Prediction Markets and Forecasting Platforms

While full integration remains future work, the methodology anticipates connection to live forecasting data as a critical enhancement. The vision is compelling: formal models grounded in collective intelligence, updating as new information emerges.

The planned approach would involve:

Semantic Matching: Model variables rarely align perfectly with forecast questions. "AI causes human extinction" might map to multiple specific forecasts about capabilities, deployment, and impacts. Developing robust matching algorithms is essential.

Temporal Alignment: Markets predict specific dates ("AGI by 2030") while models consider scenarios ("given AGI development"). Bridging these requires careful probability conditioning.

Quality Weighting: Not all forecasts are created equal. Platform reputation, forecaster track records, and market depth all affect reliability. The methodology must account for this heterogeneity.

Update Scheduling: Real-time updates would overwhelm users and computation. The system needs intelligent policies about when model updates provide value.

Platforms like Metaculus Tetlock [52] already demonstrate sophisticated conditional forecasting on AI topics. The challenge lies not in data availability but in meaningful integration that enhances rather than complicates decision-making.

With these theoretical foundations and methodological commitments established, we can now turn to the concrete implementation of AMTAIR. The next chapter demonstrates how these abstract principles translate into working software that addresses real governance challenges. The journey from theory to practice always involves surprises—some pleasant, others less so—but that's what makes it interesting.

3. AMTAIR: Design and Implementation

The moment of truth in any research project comes when elegant theories meet stubborn reality. For AMTAIR, this meant transforming the vision of automated argument extraction into working code that could handle the beautiful messiness of real AI safety arguments. Let me take you through this journey from blueprint to implementation, complete with victories, defeats, and the occasional moment of "well, that's unexpected."

3.1 System Architecture Overview

Picture, if you will, a factory for transforming arguments into models. Raw materials enter at one end—PDFs thick with jargon, blog posts mixing insight with speculation, research papers where crucial assumptions hide in footnote 47. Finished products emerge at the other end—clean network diagrams where you can trace how Assumption A leads to Catastrophe B with probability 0.3. Actually, scratch the factory metaphor. It's too clean, too industrial. This is more like archaeology meets interpretation meets mathematics. You're digging through layers of argument, trying to distinguish the load-bearing claims from rhetorical flourishes, all while preserving enough context that the formalization means something.

The pipeline consists of five main stages:

1. **Text Ingestion and Preprocessing**: Like a careful librarian, this stage catalogues incoming documents, normalizes their format, extracts metadata, and identifies the argumentative content worth processing.
2. **Argument Extraction**: The intellectual heart of the system, where large language models perform their magic, transforming prose into structured representations.
3. **Data Transformation**: The workshop where extracted arguments are refined, validated, and prepared for mathematical representation.
4. **Network Construction**: The assembly line where formal Bayesian networks are instantiated, complete with conditional probability tables.
5. **Interactive Visualization**: The showroom where complex models become accessible through thoughtful design and interactivity.

3.1.1 Five-Stage Pipeline Architecture

Let's examine each stage more closely, understanding not just what they do but why they exist as separate components.

Text Ingestion and Preprocessing handles the unglamorous but essential work of standardization. Academic PDFs, with their two-column layouts and embedded figures, differ vastly from blog posts with inline code and hyperlinks. This stage creates a uniform representation while preserving essential structure and metadata. Format normalization strips away presentation while preserving content. Metadata extraction captures authorship, publication date, and citations. Relevance filtering identifies sections containing arguments rather than literature reviews or acknowledgments. Character encoding standardization prevents those maddening replacement characters that plague text processing.

Argument Extraction represents AMTAIR's core innovation. Using a two-stage process that mirrors human reasoning, it first identifies structural relationships (what influences what) then quantifies those relationships (how likely, how strong). This separation enables targeted prompts optimized for each task, human verification between stages, and modular improvements as LLM capabilities evolve.

Data Transformation bridges the gap between textual representations and mathematical models. It parses the BayesDown syntax into structured data, validates that the resulting network forms a proper DAG, checks probability consistency, and handles missing data intelligently.

Network Construction instantiates the formal mathematical model. This involves creating nodes and edges according to extracted structure, populating conditional probability tables, initializing inference engines, and validating the complete model.

Interactive Visualization makes the complex accessible. Through thoughtful visual encoding of probabilities and relationships, progressive disclosure of detail, interactive exploration capabilities, and multiple export formats, it serves diverse stakeholder needs.

3.1.2 Design Principles

Core Design Philosophy: The architecture embodies several principles that guided countless implementation decisions:

Modularity: Each component has clear inputs, outputs, and responsibilities. This isn't just good software engineering—it enables independent improvement of components and graceful degradation when parts fail.

Validation Checkpoints: Between each stage, we validate outputs before proceeding. Bad extractions don't propagate into visualization. Malformed networks trigger re-extraction rather than cryptic errors.

Human-in-the-Loop: While pursuing automation, we recognize that human judgment remains invaluable. The architecture provides natural intervention points where experts can verify and correct.

Extensibility: New document formats, improved extraction prompts, alternative visualization libraries—the architecture accommodates growth without restructuring.

The system emphasizes transparency over black-box efficiency. Users can inspect intermediate representations, understand extraction decisions, and verify transformations. This builds trust—essential for a system handling high-stakes arguments about existential risk.

3.2 The Two-Stage Extraction Process

The heart of AMTAIR beats with a two-stage rhythm: structure, then probability. This separation, which initially seemed like an implementation detail, revealed itself as fundamental to the extraction challenge.

3.2.1 Stage 1: Structural Extraction (ArgDown)

Imagine reading a complex argument about AI risk. Your first pass likely isn't calculating exact probabilities—you're mapping the landscape. What are the key claims? How do they relate? What supports what? Stage 1 mirrors this cognitive process.

The extraction begins with pattern recognition. Natural language contains linguistic markers of causal relationships: "leads to," "results in," "depends on," "influences." The LLM, trained on vast corpora of argumentative text, recognizes these patterns and their variations.

Consider extracting from a passage like: "The development of artificial general intelligence will likely lead to rapid capability gains through recursive self-improvement. This intelligence explosion could result in systems pursuing convergent instrumental goals, potentially including resource acquisition and self-preservation. Without solved alignment, such power-seeking behavior poses existential risks to humanity."

The system identifies three key variables connected by causal relationships:

- AGI Development → Intelligence Explosion
- Intelligence Explosion → Power-Seeking Behavior
- Power-Seeking Behavior → Existential Risk

But extraction goes beyond simple pattern matching. The system must handle complex linguistic phenomena like coreference ("this," "such systems"), implicit relationships, conditional statements, and negative statements. The magic lies in prompt engineering that guides the LLM to consistent extraction while remaining flexible enough for diverse argument styles.

The output, formatted in ArgDown syntax, preserves both structure and semantics:

```
[Existential_Risk]: Threat to humanity's continued existence and flourishing.
 + [Power_Seeking_Behavior]: AI systems pursuing instrumental goals like resource acquisitio
   + [Intelligence_Explosion]: Rapid recursive self-improvement leading to superintelligence
     + [AGI_Development]: Creation of artificial general intelligence systems.
```

3.2.2 Stage 2: Probability Integration (BayesDown)

With structure established, Stage 2 adds the quantitative flesh to the qualitative bones. This stage faces a different challenge: extracting numerical beliefs from text that often expresses uncertainty in frustratingly vague terms.

The process begins by generating targeted questions based on the extracted structure. For each node, we need prior probabilities. For each child-parent relationship, we need conditional probabilities. The combinatorics can be daunting—a node with three binary parents requires 8 conditional probability values.

The system employs multiple strategies for probability extraction:

Explicit Extraction: When authors provide numerical estimates ("we assign 70% probability"), extraction is straightforward, though we must handle various formats and contexts.

Linguistic Mapping: While comprehensive validation remains future work, preliminary assessments using the methodology described above would likely reveal several patterns.

Comparative Reasoning: Statements like "more probable than not" or "at least as likely as X" provide bounds even without exact values.

Coherence Enforcement: Probabilities must sum correctly. If P(A|B) = 0.7, then P(not A|B) must equal 0.3. The syntax allows future system to detect and resolve inconsistencies.

The result is a complete BayesDown specification:

```
[Existential_Risk]: Threat to humanity's continued existence. {
  "instantiations": ["true", "false"],
  "priors": {"p(true)": "0.10", "p(false)": "0.90"},
  "posteriors": {
    "p(true|power_seeking_true)": "0.65",
    "p(true|power_seeking_false)": "0.001"
  }
}
```

3.2.3 Why Two Stages?

The separation of structure from probability isn't merely convenient—it's cognitively valid and practically essential. Let me count the ways this design decision pays dividends:

Cognitive Alignment: Humans naturally separate "what relates to what" from "how likely is it." The two-stage process mirrors this, making the system's operation intuitive and interpretable.

Error Isolation: Structural errors (missing a key variable) differ fundamentally from probability errors (estimating 0.7 instead of 0.8). Separating stages allows targeted debugging and improvement.

Modular Validation: Experts can verify structure without needing to evaluate every probability. This enables efficient human oversight at natural checkpoints.

Flexible Quantification: Different probability sources (text extraction, expert elicitation, market data) can feed into the same structure. The architecture accommodates multiple approaches to the probability challenge.

Transparency: Users can inspect ArgDown to understand what was extracted before probabilities were added. This builds trust and enables meaningful correction.

The two-stage approach also revealed an unexpected benefit: ArgDown itself became a valuable output. Researchers began using these structural extractions for qualitative analysis, even without probability quantification. Sometimes, just making argument structure explicit provides sufficient value.

3.3 Implementation Technologies

Choosing technologies for AMTAIR resembled assembling a band—each instrument needed to excel individually while harmonizing with the ensemble. The selection criteria balanced capability, maturity, interoperability, and community support.

3.3.1 Technology Stack

Selecting technologies for a project like AMTAIR involves a peculiar form of fortune-telling. You're choosing tools not just for present needs but for future possibilities you can't fully anticipate. Early decisions cascade through the implementation, creating path dependencies that only become apparent months later.

The choice of Python as the primary language was perhaps the only decision that never faced serious questioning. The ecosystem for scientific computing, the availability of sophisticated libraries, the community support—all pointed in the same direction. Yet even this "obvious" choice carried hidden implications. Python's flexibility enabled rapid prototyping but occasionally masked performance issues until they became critical.

NetworkX emerged as the natural choice for graph manipulation after brief flirtations with alternatives. Its maturity showed in countless small conveniences—algorithms I didn't have to implement, edge cases already handled, documentation for obscure functions. Pgmpy for Bayesian network operations was less obvious. Several libraries offered similar functionality, but pgmpy's API design aligned well with our extraction pipeline. The ability to construct networks incrementally, validate structure during construction, and perform inference without elaborate setup proved decisive.

The visualization challenge nearly derailed the project. Initial attempts with matplotlib produced static images that technically displayed the network but failed to convey understanding. The breakthrough came with PyVis, which leveraged vis.js to create interactive web-based visualizations. Suddenly, complex networks became explorable. Users could drag nodes to untangle

connections, click for details, adjust physics parameters to find optimal layouts. The difference between seeing and understanding turned out to be interactivity.

The final ensemble performs beautifully:

Table 3: Table 3.3.1: Overview of Tech Stack

Component	Technology	Purpose	Why This Choice
Language Models	GPT-4, Claude	Argument extraction	State-of-the-art reasoning capabilities
Network Analysis	NetworkX	Graph algorithms	Mature, comprehensive, well-documented
Probabilistic Modeling	pgmpy	Bayesian operations	Native Python, active development
Visualization	PyVis	Interactive rendering	Web-based, customizable, responsive
Data Processing	Pandas	Structured manipulation	Industry standard, powerful operations

Language Models form the cognitive core. GPT-4 and Claude demonstrate remarkable ability to understand complex arguments, recognize implicit structure, and maintain coherence across long extractions. The choice to support multiple models provides robustness and allows leveraging their complementary strengths.

NetworkX handles all graph-theoretic heavy lifting. From basic operations like cycle detection to advanced algorithms like centrality measurement, it provides a comprehensive toolkit that would take years to replicate.

pgmpy bridges the gap between graph structure and probabilistic reasoning. Its clean API design maps naturally onto our extracted representations, while its inference algorithms handle the computational complexity of Bayesian reasoning.

PyVis transforms static networks into living documents. Built on vis.js, it provides smooth physics simulations, rich interactivity, and extensive customization options—all accessible through Python.

Pandas might seem mundane compared to its companions, but it's the reliable rhythm section that keeps everything together. Its ability to reshape, merge, and transform structured data makes the complex data transformations tractable.

3.3.2 Key Algorithms

Beyond the libraries lie custom algorithms that address AMTAIR-specific challenges:

Hierarchical Parsing: The algorithm that transforms indented ArgDown text into structured data represents a small miracle of recursive descent parsing adapted for our custom syntax. It maintains parent-child relationships while handling edge cases like repeated nodes and complex dependencies.

python

```
#| label: example_use_case
#| echo: true
#| eval: true
#| fig-cap: "example use case"
#| fig-link: "https://colab.research.google.com/github/VJMeyer/submission/blob/main/AMTAIR_P
#| fig-alt: "example use case"

def parsing_argdown_bayesdown(text, current_indent=0):
    """Recursively parse indented structure maintaining relationships"""
    # Track nodes at each level for parent identification
    # Handle repeated nodes by reference
    # Validate DAG property during construction
```

Probability Completion: Real arguments rarely specify all required probabilities. Our completion algorithm uses maximum entropy principles—when uncertain, assume maximum disorder. This provides conservative estimates that can be refined with additional information.

Visual Encoding: The algorithm mapping probabilities to colors uses perceptual uniformity. The green-to-red gradient isn't linear in RGB space but follows human perception of color difference. Small details, big impact on usability.

Layout Optimization: Force-directed layouts often produce "hairballs" for complex networks. Our customized approach uses hierarchical initialization based on causal depth, then refines with physics simulation. The result: layouts that reveal structure rather than obscuring it.

3.3.3 (Expected) Performance Characteristics

Performance in a system like AMTAIR involves multiple dimensions—speed, accuracy, scalability. Let's examine what theoretical analysis and design considerations suggest about system behavior.

Computational Complexity: The extraction phase exhibits linear complexity in document length—processing twice as much text takes roughly twice as long. However, the inference phase faces exponential complexity in network connectivity. A fully connected network with n binary nodes requires O(2^n) operations for exact inference. This fundamental limitation shapes practical usage patterns.

Practical Implications: Small networks (<20 nodes) enable real-time interaction with exact inference. Medium networks (20-50 nodes) require seconds to minutes depending on connectivity. Large networks (>50 nodes) necessitate approximate methods, trading accuracy for tractability.

Very large networks push the boundaries of current methods.

The bottleneck shifts predictably: extraction remains manageable even for lengthy documents, but inference becomes challenging as models grow. This suggests a natural workflow—extract comprehensively, then focus on relevant subnetworks for detailed analysis.

Optimization Opportunities: Several strategies could improve performance: caching frequent inference queries, hierarchical decomposition of large networks, parallel processing for independent subgraphs, and progressive rendering for visualization. The modular architecture accommodates these enhancements without fundamental restructuring.

3.3.4 Deterministic vs. Probabilistic Components of the Workflow

An interesting philosophical question arises: in a system reasoning about probability, which components should themselves be probabilistic?

The current implementation draws a clear line:

Deterministic Components: All data transformations, graph algorithms, and inference calculations operate deterministically. Given the same input, they produce identical output. This provides reproducibility and debuggability—essential for building trust.

Probabilistic Components: The LLM calls for extraction introduce variability. Even with temperature set to 0, language models exhibit some randomness. Different runs might extract slightly different structures or probability estimates from the same text.

This division reflects a deeper principle: use determinism wherever possible, embrace probability where necessary. The extraction task—interpreting natural language—inherently involves uncertainty. But once we have formal representations, all subsequent operations should be predictable.

From an information-theoretic perspective, we're trying to extract maximum information from documents within computational budget constraints. Each document contains some finite amount of formalizable argument structure. Our goal is recovering as much as possible given realistic resource limits.

The two-stage extraction can be viewed as successive refinement—first recovering the higher-order bits (structure), then filling in lower-order bits (probabilities). This aligns with rate-distortion theory, where we get the most important information first.

3.4 Case Study: Rain-Sprinkler-Grass

Every field has its canonical examples—physics has spherical cows, economics has widget factories, and Bayesian networks have the rain-sprinkler-grass scenario. Despite its simplicity, this example teaches profound lessons about causal reasoning and serves as the perfect test case for AMTAIR.

3.4.1 Processing Steps

Let me walk you through how AMTAIR processes this foundational example:

The input arrives as a simple text description: "When it rains, the grass gets wet. The sprinkler also makes the grass wet. However, when it rains, we usually don't run the sprinkler."

From this prosaic description, the system performs five transformations:

1. **ArgDown Parsing**: Extract three variables (Rain, Sprinkler, Grass_Wet) and identify that rain influences both sprinkler usage and grass wetness, while the sprinkler also influences grass wetness.
2. **Question Generation**: Create probability queries: What's P(Rain)? What's P(Sprinkler|Rain)? What's P(Grass_Wet|Rain,Sprinkler) for all combinations?
3. **BayesDown Extraction**: Either extract probabilities from text or apply reasonable defaults. The "usually don't run" becomes P(Sprinkler|Rain) 0.01.
4. **Network Construction**: Build the formal Bayesian network with three nodes, three edges, and complete conditional probability tables.
5. **Visualization Rendering**: Create an interactive display where rain appears as a root cause, influencing both sprinkler and grass directly.

Each step validates its outputs before proceeding, ensuring that errors don't cascade through the pipeline.

3.4.2 Example Conversion Steps

Let's trace the actual transformations to see the pipeline in action:

Initial ArgDown Extraction:

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

The hierarchy captures that rain influences sprinkler usage—a subtle but important causal relationship that pure correlation would miss.

Generated Questions for Probability Extraction:

```
BayesDown Format Preview:
# BayesDown Representation with Placeholder Probabilities

/* This file contains BayesDown syntax with placeholder probabilities.
   Replace the placeholders with actual probability values based on the
   questions in the comments. */

    /* What is the probability for Grass_Wet=grass_wet_TRUE? */
    /* What is the probability for Grass_Wet=grass_wet_TRUE if Rain=rain_TRUE, Sprinkler=sprinkler_TRUE? */
    /* What is the probability for Grass_Wet=grass_wet_TRUE if Rain=rain_TRUE, Sprinkler=sprinkler_FALSE? */
    /* What is the probability for Grass_Wet=grass_wet_TRUE if Rain=rain_FALSE, Sprinkler=sprinkler_TRUE? */
    /* What is the probability for Grass_Wet=grass_wet_TRUE if Rain=rain_FALSE, Sprinkler=sprinkler_FALSE? */
    /* What is the probability for Grass_Wet=grass_wet_FALSE? */
    /* What is the probability for Grass_Wet=grass_wet_FALSE if Rain=rain_TRUE, Sprinkler=sprinkler_TRUE? */
    /* What is the probability for Grass_Wet=grass_wet_FALSE if Rain=rain_TRUE, Sprinkler=sprinkler_FALSE? */
    /* What is the probability for Grass_Wet=grass_wet_FALSE if Rain=rain_FALSE, Sprinkler=sprinkler_TRUE? */
```

```
    /* What is the probability for Grass_Wet=grass_wet_FALSE if Rain=rain_FALSE, Sprinkler=sprinkler_FALSE? */
    [Grass_Wet]: Concentrated moisture on, between and around the blades of grass. {"instantiations": ["grass_wet_TRUE", "grass_wet_FAL
        /* What is the probability for Rain=rain_TRUE? */
        /* What is the probability for Rain=rain_FALSE? */
        + [Rain]: Tears of angles crying high up in the skies hitting the ground. {"instantiations": ["rain_TRUE", "rain_FALSE"], "pric
        /* What is the probability for Sprinkler=sprinkler_TRUE? */
        /* What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_TRUE? */
        /* What is the probability for Sprinkler=sprinkler_TRUE if Rain=rain_FALSE? */
        /* What is the probability for Sprinkler=sprinkler_FALSE? */
        /* What is the probability for Sprinkler=sprinkler_FALSE if Rain=rain_TRUE? */
        /* What is the probability for Sprinkler=sprinkler_FALSE if Rain=rain_FALSE? */
        + [Sprinkler]: Activation of a centrifugal force based CO2 droplet distribution system. {"instantiations": ["sprinkler_TRUE", "
            /* What is the probability for Rain=rain_TRUE? */
            /* What is the probability for Rain=rain_FALSE? */
            + [Rain]
```

The system generates exactly the questions needed to fully specify the network.

Complete BayesDown Result:

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

Notice how the probabilities tell a coherent story—grass is almost certainly wet if either water source is active, almost certainly dry if neither is.

Resulting DataFrame Structure:

The transformation into tabular format enables standard data analysis tools while preserving all relationships and probabilities. Each row represents

a node with its properties, parents, children, and probability distributions.

3.4.3 Results

Table 4: Table 3.5.3: Extracted BayesDown data structure for rain-sprinkler-grass example

Title	Description	line	line_numbers	indentation	indentation_level	Parents	Children	instantiations	priors	posteriors	No_Parents	No_Children	parent_instantiations
Grass_Wet	Concentrated moisture on, between and around the blades of grass	3	[3]	0	[0]	[Rain, Sprinkler]	[]	[grass_wet_TRUE, grass_wet_FALSE]	{'p(grass_wet_TRUE)': '0.678'}	{'p(grass_wet_TRUE|sprinkler_TRUE,rain_TRUE)': '0.99', 'p(grass_wet_TRUE|sprinkler_TRUE,rain_FALSE)': '0.9', 'p(grass_wet_TRUE|sprinkler_FALSE,rain_TRUE)': '0.8', 'p(grass_wet_TRUE|sprinkler_FALSE,rain_FALSE)': '0.01'}	2	0	[[rain_TRUE, sprinkler_TRUE], [rain_FALSE, sprinkler_TRUE], [rain_TRUE, sprinkler_FALSE], [rain_FALSE, sprinkler_FALSE]]
Rain	Tears of angles crying high up in the skies hitting the ground	4	[4, 6]	2	[1, 2]	[]	[Grass_Wet, Sprinkler]	[rain_TRUE, rain_FALSE]	{'p(rain_TRUE)': '0.2', 'p(rain_FALSE)': '0.8'}		True	False	[]

Title	Description	line	line_number	indentation	indentation_level	Parents	Children	instantiations	priors	posteriors	No_Parents	No_Children	parent_instantiations
Sprinkler	Activation of a centrifugal force based CO2 droplet distribution system	5	[5]	1	[1]	[Rain]	[Grass_Wet]	[sprinkler_TRUE, sprinkler_FALSE]	{'p(sprinkler_TRUE)': '0.44838', 'p(sprinkler_FALSE)': '0.55162'}	{'p(sprinkler_TRUE|rain_TRUE)': '0.01', 'p(sprinkler_FALSE|rain_TRUE)': '0.4'}	1	1	[[rain_TRUE, rain_FALSE]]

The successfully processed rain-sprinkler-grass example demonstrates several key capabilities:

Structure Preservation: The causal relationships—including the subtle influence of rain on sprinkler usage—are correctly captured and maintained throughout processing.

Probability Coherence: All probability distributions sum to 1.0, conditional probabilities are complete, and the values tell a plausible story.

Visual Clarity: The rendered network clearly shows rain as the root cause, influencing both sprinkler and grass, while sprinkler provides an additional pathway to wet grass.

Interactive Exploration: Users can click nodes to see detailed probabilities, drag to rearrange for clarity, and explore how changing parameters affects outcomes.

Inference Capability: The system correctly calculates derived probabilities like P(Rain|Grass_Wet)—the diagnostic reasoning from effect to cause that makes Bayesian networks so powerful.

This simple example validates the basic pipeline functionality. But the real test comes with complex, real-world arguments …

Rain-Sprinkler-Grass Network Rendering

```
from IPython.display import IFrame

IFrame(src="https://singularitysmith.github.io/AMTAIR_Prototype/bayesian_network.html", width="100%", height="600px")
```

```
<IPython.lib.display.IFrame at 0x1061665d0>
```

Dynamic Html Rendering of the Rain-Sprinkler-Grass DAG with Conditional Probabilities

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

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

Carlsmith's model represents a dramatic increase in complexity—both conceptually and computationally. Where rain-sprinkler-grass has 3 nodes, Carlsmith involves 23. Where grass wetness is intuitive, "mesa-optimization" and "corrigibility" require careful thought.

3.5.1 Model Complexity

The numbers tell only part of the story:

- **23 nodes**: Each representing a substantive claim about AI development, deployment, or risk
- **29 edges**: Encoding causal relationships across technical, strategic, and societal domains
- **Multiple probability tables**: Many nodes have several parents, creating combinatorial explosion
- **Six-level causal depth**: From root causes to final catastrophe, influence propagates through multiple stages

But the conceptual complexity dwarfs the computational. Nodes like "APS-Systems" (Advanced, Planning, Strategically aware) encode specific technical hypotheses. Relationships like how "incentives to build" influence "deployment despite misalignment" require understanding of organizational behavior under competitive pressure.

This is no longer a toy problem but a serious attempt to formalize one of the most important arguments of our time.

3.5.2 Automated Extraction of the Carlsmith's Argument Structure

The extraction process began with feeding Carlsmith's paper to AMTAIR. Watching the system work felt like observing an archaeological excavation—layers of argument slowly revealed their structure.

The LLM prompts for extraction deserve special attention. Through iterative refinement, we developed prompts that guide extraction while remaining flexible:

```
#| label: prompt_template_function
#| echo: true
#| eval: true
#| fig-cap: "Prompt Template Function Definitions"
#| fig-link: "https://colab.research.google.com/github/VJMeyer/submission/blob/main/AMTAIR_Prototype/data/example_carlsmith/AMTAIR_Prot
#| fig-alt: "Prompt Template Function Definitions"
```

```
# @title 1.2.0 --- Prompt Template Function Definitions --- [prompt_template_function]

"""
BLOCK PURPOSE: Defines a flexible template system for LLM prompts used in the extraction pipeline.

This block implements two key classes:
1. PromptTemplate: A template class supporting variable substitution for dynamic prompts
2. PromptLibrary: A collection of pre-defined prompt templates for different extraction tasks

These templates are used in the ArgDown and BayesDown probability extraction
stages of the pipeline, providing consistent and well-structured prompts to the LLMs.

DEPENDENCIES: string.Template for variable substitution
OUTPUTS: PromptTemplate and PromptLibrary classes
"""
```

```
from string import Template
from typing import Dict, Optional, Union, List

class PromptTemplate:
    """Template system for LLM prompts with variable substitution"""

    def __init__(self, template: str):
        """Initialize with template string using $variable format"""
        self.template = Template(template)

    def format(self, **kwargs) -> str:
        """Substitute variables in the template"""
        return self.template.safe_substitute(**kwargs)

    @classmethod
    def from_file(cls, filepath: str) -> 'PromptTemplate':
        """Load template from a file"""
        with open(filepath, 'r') as f:
            template = f.read()
        return cls(template)

class PromptLibrary:
    """Collection of prompt templates for different extraction tasks"""

    # ArgDown extraction prompt - transforms source text into structured argument map
    ARGDOWN_EXTRACTION = PromptTemplate("""
```

```
You are participating in the AMTAIR (Automating Transformative AI Risk Modeling)
project and you are tasked with converting natural language arguments into
ArgDown syntax by extracting and formalizing causal world models from
unstructured text.
Your specific task is to extract the implicit causal model from the provided
document in structured ArgDown format.

## Epistemic Foundation & Purpose

This extraction represents one possible interpretation of the implicit causal
model in the document. Multiple extractions from the same text help reveal
patterns of convergence (where the model is clearly articulated) and
divergence (where the model contains ambiguities). This approach acknowledges
that expert texts often contain implicit rather than explicit causal models.

Your role is to reveal the causal structure already present in the author's
thinking, maintaining epistemic humility about your interpretation while
adhering strictly to the required format.

## ArgDown Format Specification

### Core Syntax

ArgDown represents causal relationships using a hierarchical structure:

1. Variables appear in square brackets with descriptive text:
   `[Variable_Name]: Description of the variable.`
```

````
2. Causal relationships use indentation (2 spaces per level) and '+' symbols:

[Effect]: Description of effect. + [Cause]: Description of cause. + [Deeper_Cause]: Description of deeper cause.

3. Causality flows from bottom (more indented) to top (less indented):
- More indented variables (causes) influence less indented variables (effects)
- The top-level variable is the ultimate effect or outcome
- Deeper indentation levels represent root causes or earlier factors

4. Each variable must include JSON metadata with possible states (instantiations):
`[Variable]: Description. {"instantiations": ["variable_STATE1", "variable_STATE2"]}`

### JSON Metadata Format

The JSON metadata must follow this exact structure:

```json
{"instantiations": ["variable_STATE1", "variable_STATE2"]}

Requirements:
* Double quotes (not single) around field names and string values
* Square brackets enclosing the instantiations array
* Comma separation between array elements
* No trailing comma after the last element
* Must be valid JSON syntax that can be parsed by standard JSON parsers
````
````

```
For binary variables (most common case):
{"instantiations": ["variable_TRUE", "variable_FALSE"]}

For multi-state variables (when clearly specified in the text):
{"instantiations": ["variable_HIGH", "variable_MEDIUM", "variable_LOW"]}

The metadata must appear on the same line as the variable definition, after the description.
## Complex Structural Patterns
### Variables Influencing Multiple Effects
The same variable can appear multiple times in different places in the hierarchy if it influences multiple effects:
[Effect1]: First effect description. {"instantiations": ["effect1_TRUE", "effect1_FALSE"]}
  + [Cause_A]: Description of cause A. {"instantiations": ["cause_a_TRUE", "cause_a_FALSE"]}

[Effect2]: Second effect description. {"instantiations": ["effect2_TRUE", "effect2_FALSE"]}
  + [Cause_A]
  + [Cause_B]: Description of cause B. {"instantiations": ["cause_b_TRUE", "cause_b_FALSE"]}

### Multiple Causes of the Same Effect
Multiple causes can influence the same effect by being listed at the same indentation level:
[Effect]: Description of effect. {"instantiations": ["effect_TRUE", "effect_FALSE"]}
  + [Cause1]: Description of first cause. {"instantiations": ["cause1_TRUE", "cause1_FALSE"]}
  + [Cause2]: Description of second cause. {"instantiations": ["cause2_TRUE", "cause2_FALSE"]}
    + [Deeper_Cause]: A cause that influences Cause2. {"instantiations": ["deeper_cause_TRUE", "deeper_cause_FALSE"]}

### Causal Chains
Causal chains are represented through multiple levels of indentation:
[Ultimate_Effect]: The final outcome. {"instantiations": ["ultimate_effect_TRUE", "ultimate_effect_FALSE"]}
```

```
  + [Intermediate_Effect]: A mediating variable. {"instantiations": ["intermediate_effect_TRUE", "intermediate_effect_FALSE"]}
    + [Root_Cause]: The initial cause. {"instantiations": ["root_cause_TRUE", "root_cause_FALSE"]}
  + [2nd_Intermediate_Effect]: A mediating variable. {"instantiations": ["intermediate_effect_TRUE", "intermediate_effect_FALSE"]}


### Common Cause of Multiple Variables
A common cause affecting multiple variables is represented by referencing the same variable in multiple places:
[Effect1]: First effect description. {"instantiations": ["effect1_TRUE", "effect1_FALSE"]}
  + [Common_Cause]: Description of common cause. {"instantiations": ["common_cause_TRUE", "common_cause_FALSE"]}

[Effect2]: Second effect description. {"instantiations": ["effect2_TRUE", "effect2_FALSE"]}
  + [Common_Cause]

## Detailed Extraction Workflow
Please follow this step-by-step process, documenting your reasoning in XML tags:
<analysis>
First, conduct a holistic analysis of the document:
1. Identify the main subject matter or domain
2. Note key concepts, variables, and factors discussed
3. Pay attention to language indicating causal relationships (causes, affects, influences, depends on, etc.)
4. Look for the ultimate outcomes or effects that are the focus of the document
5. Record your general understanding of the document's implicit causal structure
</analysis>
<variable_identification>
Next, identify and list the key variables in the causal model:
* Focus on factors that are discussed as having an influence or being influenced
* For each variable:
```

```
  * Create a descriptive name in [square_brackets]
  * Write a concise description based directly on the text
  * Determine possible states (usually binary TRUE/FALSE unless clearly specified)
* Distinguish between:
  * Outcome variables (effects the author is concerned with)
  * Intermediate variables (both causes and effects in chains)
  * Root cause variables (exogenous factors in the model)
* List all identified variables with their descriptions and possible states
</variable_identification>

<causal_structure>
Then, determine the causal relationships between variables:
* For each variable, identify what factors influence it
* Note the direction of causality (what causes what)
* Look for mediating variables in causal chains
* Identify common causes of multiple effects
* Capture feedback loops if present (though they must be represented as DAGs)
* Map out the hierarchical structure of the causal model
</causal_structure>

<format_conversion>
Now, convert your analysis into proper ArgDown format:
* Start with the ultimate outcome variables at the top level
* Place direct causes indented below with \+ symbols
* Continue with deeper causes at further indentation levels
* Add variable descriptions and instantiations metadata
* Ensure variables appearing in multiple places have consistent names
```

```
* Check that the entire structure forms a valid directed acyclic graph
</format_conversion>

<validation>

Finally, review your extraction for quality and format correctness:
1. Verify all variables have properly formatted metadata
2. Check that indentation properly represents causal direction
3. Confirm the extraction accurately reflects the document's implicit model
4. Ensure no cycles exist in the causal structure
5. Verify that variables referenced multiple times are consistent
6. Check that the extraction would be useful for subsequent analysis

</validation>


## Source Document Analysis Guidance
When analyzing the source document:
* Focus on revealing the author's own causal model, not imposing an external framework
* Maintain the author's terminology where possible
* Look for both explicit statements of causality and implicit assumptions
* Pay attention to the relative importance the author assigns to different factors
* Notice where the author expresses certainty versus uncertainty
* Consider the level of granularity appropriate to the document's own analysis

Remember that your goal is to make the implicit model explicit, not to evaluate or improve it.
The value lies in accurately representing the author's perspective, even if you might personally disagree or see limitations in their m
```

```
""")

    # BayesDown probability extraction prompt - enhances ArgDown with probability information
    BAYESDOWN_EXTRACTION = PromptTemplate("""
You are an expert in probabilistic reasoning and Bayesian networks. Your task is
to extend the provided ArgDown structure with probability information,
creating a BayesDown representation.

For each statement in the ArgDown structure, you need to:
1. Estimate prior probabilities for each possible state
2. Estimate conditional probabilities given parent states
3. Maintain the original structure and relationships

Here is the format to follow:
[Node]: Description. { "instantiations": ["node_TRUE", "node_FALSE"], "priors": { "p(node_TRUE)": "0.7", "p(node_FALSE)": "0.3" }, "pos
 [Parent]: Parent description. {...}


Here are the specific probability questions to answer:
$questions

ArgDown structure to enhance:
$argdown

Provide the complete BayesDown representation with probabilities:
""")
```

```
    @classmethod
    def get_template(cls, template_name: str) -> PromptTemplate:
        """Get a prompt template by name"""
        if hasattr(cls, template_name):
            return getattr(cls, template_name)
        else:
            raise ValueError(f"Template not found: {template_name}")
```

Prompting LLMs for ArgDown Extraction

The extraction revealed Carlsmith's elegant decomposition. At the highest level: capabilities enable power-seeking, which enables disempowerment, which constitutes catastrophe. But the details matter—deployment decisions mediated by incentives and deception, alignment difficulty influenced by multiple technical factors, corrective mechanisms that might interrupt the chain.

The ArgDown representation captured this structure:

```
# @title 1.7.0 --- Parsing ArgDown & BayesDown (.md to .csv) --- [parsing_argdown_bayesdown]

"""
BLOCK PURPOSE: Provides the core parsing functionality for transforming ArgDown
and BayesDown text representations into structured DataFrame format for further
processing.

This block implements the critical extraction pipeline described in the AMTAIR
project (see PY_TechnicalImplementation) that converts argument structures
into Bayesian networks.
The function can handle both basic ArgDown (structure-only) and
BayesDown (with probabilities).
```

```
Key steps in the parsing process:
1. Remove comments from the markdown text
2. Extract titles, descriptions, and indentation levels
3. Establish parent-child relationships based on indentation
4. Convert the structured information into a DataFrame
5. Add derived columns for network analysis

DEPENDENCIES: pandas, re, json libraries
INPUTS: Markdown text in ArgDown/BayesDown format
OUTPUTS: Structured DataFrame with node information, relationships, and properties
"""

def parse_markdown_hierarchy_fixed(markdown_text, ArgDown=False):
    """
    Parse ArgDown or BayesDown format into a structured DataFrame with parent-child relationships.

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

    Returns:
        pandas.DataFrame: Structured data with node information, relationships, and attributes
    """
    # PHASE 1: Clean and prepare the text
    clean_text = remove_comments(markdown_text)
```

```
    # PHASE 2: Extract basic information about nodes
    titles_info = extract_titles_info(clean_text)

    # PHASE 3: Determine the hierarchical relationships
    titles_with_relations = establish_relationships_fixed(titles_info, clean_text)

    # PHASE 4: Convert to structured DataFrame format
    df = convert_to_dataframe(titles_with_relations, ArgDown)

    # PHASE 5: Add derived columns for analysis
    df = add_no_parent_no_child_columns_to_df(df)
    df = add_parents_instantiation_columns_to_df(df)

    return df

def remove_comments(markdown_text):
    """
    Remove comment blocks from markdown text using regex pattern matching.

    Args:
        markdown_text (str): Text containing potential comment blocks

    Returns:
        str: Text with comment blocks removed
    """
    # Remove anything between /* and */ using regex
```

```
    return re.sub(r'/\*.*?\*/', '', markdown_text, flags=re.DOTALL)

def extract_titles_info(text):
    """
    Extract titles with their descriptions and indentation levels from markdown text.

    Args:
        text (str): Cleaned markdown text

    Returns:
        dict: Dictionary with titles as keys and dictionaries of attributes as values
    """
    lines = text.split('\n')
    titles_info = {}

    for line in lines:
        # Skip empty lines
        if not line.strip():
            continue

        # Extract title within square or angle brackets
        title_match = re.search(r'[<\[](.+?)[>\]]', line)
        if not title_match:
            continue

        title = title_match.group(1)
```

```
        # Extract description and metadata
        title_pattern_in_line = r'[<\[]' + re.escape(title) + r'[>\]]:'
        description_match = re.search(title_pattern_in_line + r'\s*(.*)', line)

        if description_match:
            full_text = description_match.group(1).strip()

            # Split description and metadata at the first "{"
            if "{" in full_text:
                split_index = full_text.find("{")
                description = full_text[:split_index].strip()
                metadata = full_text[split_index:].strip()
            else:
                # Keep the entire description and no metadata
                description = full_text
                metadata = ''  # Initialize as empty string
        else:
            description = ''
            metadata = ''  # Ensure metadata is initialized

        # Calculate indentation level based on spaces before + or - symbol
        indentation = 0
        if '+' in line:
            symbol_index = line.find('+')
            # Count spaces before the '+' symbol
            i = symbol_index - 1
            while i >= 0 and line[i] == ' ':
```

```
            indentation += 1
            i -= 1
    elif '-' in line:
        symbol_index = line.find('-')
        # Count spaces before the '-' symbol
        i = symbol_index - 1
        while i >= 0 and line[i] == ' ':
            indentation += 1
            i -= 1

    # If neither symbol exists, indentation remains 0

    if title in titles_info:
        # Only update description if it's currently empty and we found a new one
        if not titles_info[title]['description'] and description:
            titles_info[title]['description'] = description

        # Store all indentation levels for this title
        titles_info[title]['indentation_levels'].append(indentation)

        # Keep max indentation for backward compatibility
        if indentation > titles_info[title]['indentation']:
            titles_info[title]['indentation'] = indentation

        # Do NOT update metadata here - keep the original metadata
    else:
        # First time seeing this title, create a new entry
```

```
            titles_info[title] = {
                'description': description,
                'indentation': indentation,
                'indentation_levels': [indentation],  # Initialize with first indentation level
                'parents': [],
                'children': [],
                'line': None,
                'line_numbers': [],  # Initialize an empty list for all occurrences
                'metadata': metadata  # Set metadata explicitly from what we found
            }

    return titles_info

def establish_relationships_fixed(titles_info, text):
    """
    Establish parent-child relationships between titles using BayesDown
    indentation rules.

    In BayesDown syntax:
    - More indented nodes (with + symbol) are PARENTS of less indented nodes
    - The relationship reads as "Effect is caused by Cause" (Effect + Cause)
    - This aligns with how Bayesian networks represent causality

    Args:
        titles_info (dict): Dictionary with information about titles
        text (str): Original markdown text (for identifying line numbers)
```

```
    Returns:
        dict: Updated dictionary with parent-child relationships
    """
    lines = text.split('\n')

    # Dictionary to store line numbers for each title occurrence
    title_occurrences = {}

    # Record line number for each title (including multiple occurrences)
    line_number = 0
    for line in lines:
        if not line.strip():
            line_number += 1
            continue

        title_match = re.search(r'[<\[](.+?)[>\]]', line)
        if not title_match:
            line_number += 1
            continue

        title = title_match.group(1)

        # Store all occurrences of each title with their line numbers
        if title not in title_occurrences:
            title_occurrences[title] = []
        title_occurrences[title].append(line_number)
```

```
        # Store all line numbers where this title appears
        if 'line_numbers' not in titles_info[title]:
            titles_info[title]['line_numbers'] = []
        titles_info[title]['line_numbers'].append(line_number)

        # For backward compatibility, keep the first occurrence in 'line'
        if titles_info[title]['line'] is None:
            titles_info[title]['line'] = line_number

        line_number += 1

    # Create an ordered list of all title occurrences with their line numbers
    all_occurrences = []
    for title, occurrences in title_occurrences.items():
        for line_num in occurrences:
            all_occurrences.append((title, line_num))

    # Sort occurrences by line number
    all_occurrences.sort(key=lambda x: x[1])

    # Get indentation for each occurrence
    occurrence_indents = {}
    for title, line_num in all_occurrences:
        for line in lines[line_num:line_num+1]:  # Only check the current line
            indent = 0
            if '+' in line:
                symbol_index = line.find('+')
```

```
            # Count spaces before the '+' symbol
            j = symbol_index - 1
            while j >= 0 and line[j] == ' ':
                indent += 1
                j -= 1
        elif '-' in line:
            symbol_index = line.find('-')
            # Count spaces before the '-' symbol
            j = symbol_index - 1
            while j >= 0 and line[j] == ' ':
                indent += 1
                j -= 1
        occurrence_indents[(title, line_num)] = indent

# Enhanced backward pass for correct parent-child relationships
for i, (title, line_num) in enumerate(all_occurrences):
    current_indent = occurrence_indents[(title, line_num)]

    # Skip root nodes (indentation 0) for processing
    if current_indent == 0:
        continue

    # Look for the immediately preceding node with lower indentation
    j = i - 1
    while j >= 0:
        prev_title, prev_line = all_occurrences[j]
        prev_indent = occurrence_indents[(prev_title, prev_line)]
```

```
            # If we find a node with less indentation, it's a child of current node
            if prev_indent < current_indent:
                # In BayesDown:
                # More indented node is a parent (cause) of less indented node (effect)
                if title not in titles_info[prev_title]['parents']:
                    titles_info[prev_title]['parents'].append(title)
                if prev_title not in titles_info[title]['children']:
                    titles_info[title]['children'].append(prev_title)

                # Only need to find the immediate child
                # (closest preceding node with lower indentation)
                break

            j -= 1

    return titles_info

def convert_to_dataframe(titles_info, ArgDown):
    """
    Convert the titles information dictionary to a pandas DataFrame.

    Args:
        titles_info (dict): Dictionary with information about titles
        ArgDown (bool): If True, extract only structural information without probabilities

    Returns:
```

```
    pandas.DataFrame: Structured data with node information and relationships
"""
if ArgDown == True:
    # For ArgDown, exclude probability columns
    df = pd.DataFrame(columns=['Title', 'Description', 'line', 'line_numbers', 'indentation',
                           'indentation_levels', 'Parents', 'Children', 'instantiations'])
else:
    # For BayesDown, include probability columns
    df = pd.DataFrame(columns=['Title', 'Description', 'line', 'line_numbers', 'indentation',
                           'indentation_levels', 'Parents', 'Children', 'instantiations',
                           'priors', 'posteriors'])

for title, info in titles_info.items():
    # Parse the metadata JSON string into a Python dictionary
    if 'metadata' in info and info['metadata']:
        try:
            # Only try to parse if metadata is not empty
            if info['metadata'].strip():
                jsonMetadata = json.loads(info['metadata'])
                if ArgDown == True:
                    # Create the row dictionary with instantiations as
                    # metadata only, no probabilities yet
                    row = {
                        'Title': title,
                        'Description': info.get('description', ''),
                        'line': info.get('line',''),
                        'line_numbers': info.get('line_numbers', []),
```

```
                    'indentation': info.get('indentation',''),
                    'indentation_levels': info.get('indentation_levels', []),
                    'Parents': info.get('parents', []),
                    'Children': info.get('children', []),
                    # Extract specific metadata fields,
                    # defaulting to empty if not present
                    'instantiations': jsonMetadata.get('instantiations', []),
                }
            else:
                # Create dict with probabilities for BayesDown
                row = {
                    'Title': title,
                    'Description': info.get('description', ''),
                    'line': info.get('line',''),
                    'line_numbers': info.get('line_numbers', []),
                    'indentation': info.get('indentation',''),
                    'indentation_levels': info.get('indentation_levels', []),
                    'Parents': info.get('parents', []),
                    'Children': info.get('children', []),
                    # Extract specific metadata fields, defaulting to empty if not present
                    'instantiations': jsonMetadata.get('instantiations', []),
                    'priors': jsonMetadata.get('priors', {}),
                    'posteriors': jsonMetadata.get('posteriors', {})
                }
        else:
            # Empty metadata case
            row = {
```

```
                    'Title': title,
                    'Description': info.get('description', ''),
                    'line': info.get('line',''),
                    'line_numbers': info.get('line_numbers', []),
                    'indentation': info.get('indentation',''),
                    'indentation_levels': info.get('indentation_levels', []),
                    'Parents': info.get('parents', []),
                    'Children': info.get('children', []),
                    'instantiations': [],
                    'priors': {},
                    'posteriors': {}
                }
            except json.JSONDecodeError:
                # Handle case where metadata isn't valid JSON
                row = {
                    'Title': title,
                    'Description': info.get('description', ''),
                    'line': info.get('line',''),
                    'line_numbers': info.get('line_numbers', []),
                    'indentation': info.get('indentation',''),
                    'indentation_levels': info.get('indentation_levels', []),
                    'Parents': info.get('parents', []),
                    'Children': info.get('children', []),
                    'instantiations': [],
                    'priors': {},
                    'posteriors': {}
                }
```

```
        else:
            # Handle case where metadata field doesn't exist or is empty
            row = {
                'Title': title,
                'Description': info.get('description', ''),
                'line': info.get('line',''),
                'line_numbers': info.get('line_numbers', []),
                'indentation': info.get('indentation',''),
                'indentation_levels': info.get('indentation_levels', []),
                'Parents': info.get('parents', []),
                'Children': info.get('children', []),
                'instantiations': [],
                'priors': {},
                'posteriors': {}
            }

        # Add the row to the DataFrame
        df.loc[len(df)] = row

    return df

def add_no_parent_no_child_columns_to_df(dataframe):
    """
    Add No_Parent and No_Children boolean columns to the DataFrame to
    identify root and leaf nodes.

    Args:
```

```
        dataframe (pandas.DataFrame): The DataFrame to enhance

    Returns:
        pandas.DataFrame: Enhanced DataFrame with additional boolean columns
    """
    no_parent = []
    no_children = []

    for _, row in dataframe.iterrows():
        no_parent.append(not row['Parents'])  # True if Parents list is empty
        no_children.append(not row['Children'])  # True if Children list is empty

    dataframe['No_Parent'] = no_parent
    dataframe['No_Children'] = no_children

    return dataframe

def add_parents_instantiation_columns_to_df(dataframe):
    """
    Add all possible instantiations of parents as a list of lists column
    to the DataFrame.
    This is crucial for generating conditional probability tables.

    Args:
        dataframe (pandas.DataFrame): The DataFrame to enhance

    Returns:
```

```
    pandas.DataFrame: Enhanced DataFrame with parent_instantiations column
"""
# Create a new column to store parent instantiations
parent_instantiations = []

# Iterate through each row in the dataframe
for _, row in dataframe.iterrows():
    parents = row['Parents']
    parent_insts = []

    # For each parent, find its instantiations and add to the list
    for parent in parents:
        # Find the row where Title matches the parent
        parent_row = dataframe[dataframe['Title'] == parent]

        # If parent found in the dataframe
        if not parent_row.empty:
            # Get the instantiations of this parent
            parent_instantiation = parent_row['instantiations'].iloc[0]
            parent_insts.append(parent_instantiation)

    # Add the list of parent instantiations to our new column
    parent_instantiations.append(parent_insts)

# Add the new column to the dataframe
dataframe['parent_instantiations'] = parent_instantiations
```

```
    return dataframe
```

The structure revealed insights. "Misaligned_Power_Seeking" emerged as a critical hub, influenced by multiple factors and influencing multiple outcomes. The pathway from incentives through deployment to risk became explicit.

3.5.3 From ArgDown to BayesDown in Carlsmith's Model

Adding probabilities to Carlsmith's structure presented unique challenges. Unlike rain-sprinkler probabilities that have intuitive values, what's the probability of "mesa-optimization" or "deceptive alignment"?

The system generated over 100 probability questions for the full model.

Each question targets a specific parameter needed for the Bayesian network. The conditional structure reflects Carlsmith's argument—deployment depends on both incentives (external pressure) and deception (hidden misalignment).

The LLM extraction drew on Carlsmith's explicit estimates where available and inferred reasonable values elsewhere. The result captured both the structure and Carlsmith's quantitative risk assessment:

```
[Deployment_Decisions]: Decisions to deploy potentially misaligned AI systems. {
  "instantiations": ["deployment_decisions_DEPLOY", "deployment_decisions_WITHHOLD"],
  "priors": {
    "p(deployment_decisions_DEPLOY)": "0.70",
    "p(deployment_decisions_WITHHOLD)": "0.30"
  },
  "posteriors": {
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_TRUE)": "0.90",
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_STRONG, deception_by_ai_FALSE)": "0.75",
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_TRUE)": "0.60",
    "p(deployment_decisions_DEPLOY|incentives_to_build_aps_WEAK, deception_by_ai_FALSE)": "0.30"
  }
}
```

This node has two possible states (DEPLOY or WITHHOLD), prior probabilities for each state, and conditional probabilities based on different combinations of its parent variables ("Incentives_To_Build_APS" and "Deception_By_AI"). The probabilities tell a plausible story: deployment becomes more likely with stronger incentives and successful deception, but even without deception, strong incentives create substantial deployment probability.

Along with these questions the following prompt is sent to the LLM:

```
You are an expert in probabilistic reasoning and Bayesian networks. Your task is
to extend the provided ArgDown structure with probability information,
creating a BayesDown representation.

For each statement in the ArgDown structure, you need to:
1. Estimate prior probabilities for each possible state
2. Estimate conditional probabilities given parent states
3. Maintain the original structure and relationships

Here is the format to follow:
[Node]: Description. { "instantiations": ["node_TRUE", "node_FALSE"], "priors": { "p(node_TRUE)": "0.7", "p(node_FALSE)": "0.3" }, "pos
 [Parent]: Parent description. {...}


Here are the specific probability questions to answer:
$questions

ArgDown structure to enhance:
$argdown

Provide the complete BayesDown representation with probabilities:
```

Example BayesDown Excerpt from the Carlsmith model

```
#| label: json_carlsmith_excerpt
#| echo: true
#| eval: true
#| fig-cap: "Example BayesDown Excerpt from the Carlsmith model"
#| fig-link: "https://colab.research.google.com/github/VJMeyer/submission/blob/main/AMTAIR_Prototype/data/example_carlsmith/AMTAIR_Prot
#| fig-alt: "Example BayesDown Excerpt from the Carlsmith model"



[Existential_Catastrophe]: The destruction of humanity's long-term potential due to AI systems we've lost control over. {
  "instantiations": ["existential_catastrophe_TRUE", "existential_catastrophe_FALSE"],
  "priors": {"p(existential_catastrophe_TRUE)": "0.05", "p(existential_catastrophe_FALSE)": "0.95"},
  "posteriors": {
    "p(existential_catastrophe_TRUE|human_disempowerment_TRUE)": "0.95",
    "p(existential_catastrophe_TRUE|human_disempowerment_FALSE)": "0.0"
  }
}
 + [Human_Disempowerment]: Permanent and collective disempowerment of humanity relative to AI systems. {
   "instantiations": ["human_disempowerment_TRUE", "human_disempowerment_FALSE"],
   "priors": {"p(human_disempowerment_TRUE)": "0.208", "p(human_disempowerment_FALSE)": "0.792"},
   "posteriors": {
     "p(human_disempowerment_TRUE|scale_of_power_seeking_TRUE)": "1.0",
     "p(human_disempowerment_TRUE|scale_of_power_seeking_FALSE)": "0.0"
   }
 }
```

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

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

3.5.4 Practically Meaningful BayesDown

The BayesDown representation achieves something remarkable: it bridges the chasm between Carlsmith's nuanced prose and mathematical formalism without losing the essence of either.

Consider what this bridge enables:

For Technical Researchers: The formal structure makes assumptions explicit. Is power-seeking really independent of capability level given strategic awareness? The model forces clarity.

For Policymakers: Probabilities attached to comprehensible descriptions provide actionable intelligence. "70% chance of deployment despite misalignment" translates better than abstract concerns.

For Strategic Analysts: The network structure reveals intervention points. Which nodes, if changed, most affect the final outcome? Where should we focus effort?

The hybrid nature—natural language plus formal structure plus probabilities—serves each audience while enabling communication between them. A policymaker can understand "deployment decisions" without probability theory. A researcher can analyze the mathematical model without losing sight of what the variables mean.

This isn't just convenient—it's essential for coordination. When different communities can refer to the same model but engage with it at their appropriate level of technical detail, we create common ground for productive disagreement and collaborative problem-solving.

3.5.5 Interactive Visualization and Exploration

The moment when Carlsmith's model first rendered as an interactive network felt like putting on glasses after years of squinting. Suddenly, the complex web of relationships became navigable.

The visualization system employs multiple visual channels simultaneously:

Color Coding: Nodes shift from deep red (low probability) through yellow to bright green (high probability). At a glance, you see which factors Carlsmith considers likely versus speculative.

Border Styling: Blue borders mark root causes (like "Incentives_To_Build"), purple indicates intermediate nodes, magenta highlights final outcomes. The visual grammar guides the eye through causal flow.

Layout Algorithm: Initial placement uses causal depth—root causes at bottom, final outcomes at top. Physics simulation then refines positions to minimize edge crossings while preserving hierarchical structure.

Progressive Disclosure: Hovering reveals probability summaries. Clicking opens detailed conditional probability tables. Dragging allows custom arrangement. Each interaction level serves different analytical needs.

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

```
# @title 4.4.0 --- Main Visualization Function --- [main_visualization_function]

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

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

        # Process probability information
        priors = get_priors(row)
        instantiations = get_instantiations(row)

        # Add node with base information
        G.add_node(
            title,
            description=description,
            priors=priors,
            instantiations=instantiations,
            posteriors=get_posteriors(row)
        )

    # Add edges
    for idx, row in df.iterrows():
        child = row['Title']
        parents = get_parents(row)

        # Add edges from each parent to this child
        for parent in parents:
            if parent in G.nodes():
                G.add_edge(parent, child)
```

```
# Classify nodes based on network structure
classify_nodes(G)

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

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

# Add the graph to the network
net.from_nx(G)

# Enhance node appearance with probability information and classification
for node in net.nodes:
    node_id = node['id']
    node_data = G.nodes[node_id]

    # Get node type and set border color
    node_type = node_data.get('node_type', 'unknown')
    border_color = get_border_color(node_type)

    # Get probability information
    priors = node_data.get('priors', {})
    true_prob = priors.get('true_prob', 0.5) if priors else 0.5
```

```
        # Get proper state names
        instantiations = node_data.get('instantiations', ["TRUE", "FALSE"])
        true_state = instantiations[0] if len(instantiations) > 0 else "TRUE"
        false_state = instantiations[1] if len(instantiations) > 1 else "FALSE"

        # Create background color based on probability
        background_color = get_probability_color(priors)

        # Create tooltip with probability information
        tooltip = create_tooltip(node_id, node_data)

        # Create a simpler node label with probability
        simple_label = f"{node_id}\np={true_prob:.2f}"

        # Store expanded content as a node attribute for use in click handler
        node_data['expanded_content'] = create_expanded_content(node_id, node_data)

        # Set node attributes
        node['title'] = tooltip  # Tooltip HTML
        node['label'] = simple_label  # Simple text label
        node['shape'] = 'box'
        node['color'] = {
            'background': background_color,
            'border': border_color,
            'highlight': {
                'background': background_color,
                'border': border_color
```

```
        }
    }

# Set up the click handler with proper data
setup_data = {
    'nodes_data': {node_id: {
        'expanded_content': json.dumps(G.nodes[node_id].get('expanded_content', '')),
        'description': G.nodes[node_id].get('description', ''),
        'priors': G.nodes[node_id].get('priors', {}),
        'posteriors': G.nodes[node_id].get('posteriors', {})
    } for node_id in G.nodes()}
}

# Add custom click handling JavaScript
click_js = """
// Store node data for click handling
var nodesData = %s;

// Add event listener for node clicks
network.on("click", function(params) {
    if (params.nodes.length > 0) {
        var nodeId = params.nodes[0];
        var nodeInfo = nodesData[nodeId];

        if (nodeInfo) {
            // Create a modal popup for expanded content
            var modal = document.createElement('div');
```

```
            modal.style.position = 'fixed';
            modal.style.left = '50%%';
            modal.style.top = '50%%';
            modal.style.transform = 'translate(-50%%, -50%%)';
            modal.style.backgroundColor = 'white';
            modal.style.padding = '20px';
            modal.style.borderRadius = '5px';
            modal.style.boxShadow = '0 0 10px rgba(0,0,0,0.5)';
            modal.style.zIndex = '1000';
            modal.style.maxWidth = '80%%';
            modal.style.maxHeight = '80%%';
            modal.style.overflow = 'auto';

            // Parse the JSON string back to HTML content
            try {
                var expandedContent = JSON.parse(nodeInfo.expanded_content);
                modal.innerHTML = expandedContent;
            } catch (e) {
                modal.innerHTML = 'Error displaying content: ' + e.message;
            }

            // Add close button
            var closeBtn = document.createElement('button');
            closeBtn.innerHTML = 'Close';
            closeBtn.style.marginTop = '10px';
            closeBtn.style.padding = '5px 10px';
            closeBtn.style.cursor = 'pointer';
```

```
            closeBtn.onclick = function() {
                document.body.removeChild(modal);
            };
            modal.appendChild(closeBtn);

            // Add modal to body
            document.body.appendChild(modal);
        }
    }
});
""" % json.dumps(setup_data['nodes_data'])

# Save the graph to HTML
html_file = "bayesian_network.html"
net.save_graph(html_file)

# Inject custom click handling into HTML
try:
    with open(html_file, "r") as f:
        html_content = f.read()

    # Insert click handling script before the closing body tag
    html_content = html_content.replace('</body>', f'<script>{click_js}</script></body>')

    # Write back the modified HTML
    with open(html_file, "w") as f:
        f.write(html_content)
```

```
        return HTML(html_content)
    except Exception as e:
        return HTML(f"<p>Error rendering HTML: {str(e)}</p>"
        + "<p>The network visualization has been saved to '{html_file}'</p>")
```

The resulting visualization transforms abstract relationships into tangible understanding. Users report "aha" moments when exploring—suddenly seeing how technical factors compound into strategic risks, or identifying previously unnoticed bottlenecks in the causal chain.

This visualization reveals several structural insights:

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

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

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

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

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

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

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

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

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

3.5.6 Validation Against Original (From the MTAIR Project)

Validating AMTAIR's extraction required careful comparison with expert judgment. While comprehensive benchmarking remains future work, preliminary validation efforts provide encouraging signals.

Manual Baseline Creation: Johannes Meyer and Jelena Meyer, independently extracted ArgDown and BayesDown representations from Carlsmith's paper and Bucknall and Dori-Hacohen's. This created ground truth accounting for legitimate interpretive variation—experts might reasonably disagree on some structural choices or probability estimates.

Structural Comparison: Comparing extracted causal structures revealed high agreement on core relationships. AMTAIR consistently identified the main causal chain from capabilities through deployment to catastrophe. Some variation appeared in handling of auxiliary factors—where one expert might include a minor influence, another might omit it for simplicity.

Probability Assessment: Probability extraction showed greater variation, reflecting inherent ambiguity in translating qualitative language. When Carlsmith writes "likely," different readers might reasonably interpret this as 0.7, 0.75, or 0.8. AMTAIR's extractions fell within the range of expert interpretations, suggesting successful capture of intended meaning even if not identical numbers.

Semantic Preservation: Most importantly, the formal models preserved the essential insights of Carlsmith's argument. The critical role of deployment decisions, the compound nature of risk, the importance of technical and strategic factors—all emerged clearly in the extracted representations.

An ideal validation protocol would expand this approach:

1. Multiple expert extractors working independently
2. Systematic comparison of structural and quantitative agreement
3. Analysis of where and why extractions diverge
4. Testing whether different extractions lead to different policy conclusions
5. Iterative refinement based on identified failure modes

The goal isn't perfect agreement—even human experts disagree. Rather, we seek extractions good enough to support meaningful analysis while acknowledging their limitations.

3.6 Validation Methodology

Building trust in automated extraction requires more than anecdotal success. We need systematic validation that honestly assesses both capabilities and limitations.

3.6.1 Ground Truth Construction

Creating ground truth for argument extraction poses unique challenges. Unlike named entity recognition or sentiment analysis, argument structure lacks universal standards. What constitutes the "correct" extraction from a complex text?

An ideal validation approach would embrace this inherent subjectivity:

Expert Selection: Recruit 5-10 domain experts with demonstrated expertise in both AI safety and formal modeling. Diversity matters—include technical researchers, policy analysts, and those with mixed backgrounds.

Extraction Protocol: Provide standardized training on ArgDown/BayesDown syntax while allowing flexibility in interpretation. Experts work independently to avoid anchoring bias, documenting their reasoning process alongside final extractions.

Consensus Building: Through structured discussion, identify areas of convergence (likely core argument structure) versus legitimate disagreement (interpretive choices, granularity decisions). This distinguishes system errors from inherent ambiguity.

Quality Metrics: Rather than binary correct/incorrect judgments, assess:

- Structural similarity (graph edit distance)
- Probability distribution overlap (KL divergence)
- Semantic preservation (expert ratings)
- Downstream task performance (policy analysis agreement)

The resulting dataset would capture not a single "truth" but a distribution of reasonable interpretations against which to evaluate automated extraction.

3.6.2 Evaluation Metrics

Evaluating argument extraction requires metrics that capture multiple dimensions of quality:

Structural Fidelity:

- Node identification: What fraction of expert-identified variables does the system extract?
- Edge accuracy: Are causal relationships preserved?
- Hierarchy preservation: Does the system maintain argument levels?

Probability Calibration:

- Explicit extraction: When sources state probabilities, how accurately are they captured?
- Linguistic mapping: Do qualitative expressions translate to reasonable probabilities?
- Coherence: Are probability distributions properly normalized?

Semantic Quality:

- Description accuracy: Do extracted descriptions preserve original meaning?
- Terminology preservation: Does the system maintain author's vocabulary?
- Context retention: Is sufficient information preserved for interpretation?

Functional Validity:

- Inference agreement: Do extracted models support similar conclusions?
- Sensitivity preservation: Are critical parameters identified as influential?
- Policy robustness: Do different extractions suggest similar interventions?

These metrics acknowledge that perfect extraction is neither expected nor necessary. The goal is extraction sufficient for practical use while maintaining transparency about limitations.

3.6.3 Results Summary

While comprehensive validation remains future work, preliminary assessments using the methodology described above would likely reveal several patterns:

Expected Strengths: Automated extraction should excel at identifying explicit causal claims, preserving hierarchical argument structure, and extracting stated probabilities. The two-stage approach likely improves quality by allowing focused optimization for each task.

Anticipated Challenges: Implicit reasoning, complex conditionals, and ambiguous quantifiers would pose greater challenges. Coreference resolution across long documents and maintaining consistency in large models would require continued refinement.

Practical Utility Threshold: Even with imperfect extraction, the system could provide value if it achieves perhaps 70-80% structural accuracy and captures probability estimates within reasonable ranges. This level of performance would enable rapid initial modeling that experts could refine, dramatically reducing the time from argument to formal model.

The validation framework itself represents a contribution—establishing systematic methods for assessing argument extraction quality as this research area develops.

3.6.4 Error Analysis

Understanding failure modes guides both appropriate use and future improvements:

Implicit Assumptions: Authors often leave critical assumptions unstated, relying on shared background knowledge. When an AI safety researcher writes about "alignment," they assume readers understand the technical concept. The system must either extract these implicit elements or flag their absence.

Complex Conditionals: Natural language expresses conditionality in myriad ways. "If we achieve alignment (which seems unlikely without major theoretical breakthroughs), then deployment might be safe (assuming robust verification)." Parsing nested, qualified conditionals challenges current methods.

Ambiguous Quantifiers: The word "significant" might mean 10% in one context, 60% in another. Without calibration to author-specific usage or domain conventions, probability extraction remains approximate.

Coreference Challenges: Academic writing loves pronouns and indirect references. When "this approach" appears three paragraphs after introducing multiple approaches, identifying the correct referent requires sophisticated discourse understanding.

These limitations don't invalidate the approach but rather define its boundaries. Users who understand these constraints can work within them, leveraging automation's strengths while compensating for its weaknesses.

3.6.5 Independent Manual Extraction Validation

To establish ground truth for evaluating AMTAIR's extraction quality, I obtained independent manual extractions from domain experts. Johannes Meyer and Jelena Meyer[14], both experienced in formal logic and argument analysis, independently extracted ArgDown and BayesDown representations from Bucknall and Dori-Hacohen's "Current and Near-Term AI as a Potential Existential Risk Factor" Bucknall and Dori-Hacohen [8]. This paper, which examines how near-term AI systems might contribute to existential risks through various causal pathways, provides an ideal test case due to its explicit discussion of multiple risk factors and their interdependencies.

The manual extraction process revealed patterns consistent with theoretical expectations from the argument mining literature Khartabil et al. [30]. Both extractors identified remarkably similar causal structures—the core nodes representing existential risk factors (unaligned AGI, nuclear conflict, biological risks, environmental catastrophe) and their relationships to near-term AI capabilities showed near-perfect agreement. This structural convergence aligns with findings from Anderson Anderson [1] that expert annotators tend to agree on primary argumentative relationships even when working independently.

However, the probability quantification phase exhibited substantially higher variance, corroborating established challenges in eliciting subjective probabilities from text. When extracting conditional probabilities for relationships like P(Nuclear_Conflict | Compromised_Political_Decision_Making), the two extractors' estimates differed by as much as 30 percentage points. This variance reflects the fundamental ambiguity Pollock Pollock [44] identified in mapping natural language uncertainty expressions to numerical values—when Bucknall and Dori-Hacohen write that AI "may intensify cyber warfare," reasonable interpreters might assign probabilities anywhere from 0.4 to 0.7.

The extraction revealed a hierarchical structure with [Existential_Risk] as the root node, influenced by both direct AI risks (unaligned AGI) and indirect pathways where near-term AI acts as an intermediate risk factor. The extractors consistently identified four main causal mechanisms: state-to-state relations (arms race dynamics), corporate power concentration, stable repressive regimes, and compromised political decision-making. This structural clarity demonstrates that despite quantitative uncertainty, the qualitative causal model remains extractable with high fidelity.

Interestingly, both manual extractors struggled with the same ambiguities that challenge automated systems, which could be an indication about convergence on the underlying level of information contained in the source. The relationship between social media recommender systems and various risk factors appeared multiple times in the text with slightly different framings, requiring judgment calls about whether these represented single or multiple causal relationships. This observation supports the design decision to maintain human oversight in AMTAIR's extraction pipeline—certain interpretive choices require domain knowledge and contextual under-

[14] I am extremely grateful for their help, support and the invaluable contribution. As lead engineer I had had the nagging suspicion that, maybe I had "hardcoded" by own intuitions into the system (through choices in the setup, system prompt, source selection etc.). I am relieved to let go of this concern and hope that future, large scale work confirms the potential for objectivity and convergence.

standing that neither human nor machine extractors can make with complete confidence in isolation.

The manual extraction exercise validates AMTAIR's two-stage approach. The high agreement on structure (ArgDown) combined with high variance in probabilities (BayesDown) empirically confirms that separating these extraction tasks addresses genuine cognitive and epistemological differences. As predicted by the causal structure learning literature Heinze-Deml, Maathuis, and Meinshausen [26] Squires and Uhler [50], identifying "what causes what" represents a different inferential challenge than quantifying "how likely" those causal relationships are.

This validation also illuminates the value proposition of automated extraction. While human experts required 4-6 hours each to complete their extractions, AMTAIR processed the same document in under two minutes. Even if automated extraction only achieves 80% of human accuracy, the 100x speed improvement enables analyzing entire literatures rather than individual papers. The manual baseline suggests that perfect extraction may be impossible even for humans—but good-enough extraction at scale can still transform how we synthesize complex arguments about AI risk.

3.7 Policy Evaluation Capabilities

The ultimate test of a model isn't its elegance but its utility. Can AMTAIR's extracted models actually inform governance decisions? This section demonstrates how formal models enable systematic policy analysis.

3.7.1 Intervention Representation

Representing policy interventions in Bayesian networks requires translating governance mechanisms into parameter modifications. Pearl's do-calculus provides the mathematical framework, but the practical challenge lies in meaningful translation.

An ideal implementation would support several intervention types:

Parameter Modification: Policies often change probabilities. Safety requirements might reduce P(deployment|misaligned) from 0.7 to 0.2 by making unsafe deployment legally prohibited or reputationally costly.

Structural Interventions: Some policies add new causal pathways. Introducing mandatory review boards creates new nodes and edges representing oversight mechanisms.

Uncertainty Modeling: Policy effectiveness is itself uncertain. Rather than assuming perfect implementation, represent ranges: P(deployment|misaligned) might become [0.1, 0.3] depending on enforcement.

Multi-Level Effects: Policies influence multiple levels simultaneously. Compute governance affects technical development, corporate behavior, and international competition.

The system would translate high-level policy descriptions into specific network modifications, enabling rigorous counterfactual analysis of intervention effects.

3.7.2 Example: Deployment Governance

Let's trace how a specific policy—mandatory safety certification before deployment—might be evaluated:

Baseline Model: In Carlsmith's original model, P(deployment|misaligned) = 0.7, reflecting competitive pressures overwhelming safety concerns.

Policy Specification: Safety certification requires demonstrating alignment properties before deployment authorization. Based on similar regulations in other domains, we might estimate 80-90% effectiveness.

Parameter Update: The modified model sets P(deployment|misaligned) = 0.1-0.2, representing the residual probability of circumvention or regulatory capture.

Downstream Effects:

- Reduced deployment of misaligned systems
- Lower probability of power-seeking manifestation
- Decreased existential risk from ~5% to ~1.2%

Sensitivity Analysis: How robust is this conclusion? Varying certification effectiveness, enforcement probability, and other parameters reveals which assumptions critically affect the outcome.

This example illustrates policy evaluation's value: moving from vague claims ("regulation would help") to quantitative assessments ("this specific intervention might reduce risk by 75%±15%").

3.7.3 Robustness Analysis

Good policies work across scenarios. AMTAIR enables testing interventions against multiple worldviews, parameter ranges, and structural variations.

Cross-Model Testing: Extract multiple expert models and evaluate the same policy in each. If an intervention reduces risk in Carlsmith's model but increases it in Christiano's, we've identified a critical dependency.

Parameter Sensitivity: Which uncertainties most affect policy effectiveness? If the intervention only works for P(alignment_difficulty) < 0.3, and experts disagree whether it's 0.2 or 0.4, we need more research before implementing.

Structural Uncertainty: Some disagreements concern model structure itself. Does capability advancement directly influence misalignment risk, or only indirectly through deployment pressures? Test policies under both structures.

Confidence Bounds: Rather than point estimates, compute ranges. "This policy reduces risk by 40-80%" honestly represents uncertainty while still providing actionable guidance.

The goal isn't eliminating uncertainty but making decisions despite it. Robustness analysis reveals which policies work across uncertainties versus those requiring specific assumptions.

3.8 Interactive Visualization Design

A Bayesian network without good visualization is like a symphony without performers—all potential, no impact. The visualization system transforms mathematical abstractions into intuitive understanding.

3.8.1 Visual Encoding Strategy

Every visual element carries information:

Color: The probability spectrum from red (low) through yellow to green (high) provides immediate gestalt understanding. Pre-attentive processing—the brain's ability to process certain visual features without conscious attention—makes patterns jump out.

Borders: Node type encoding (blue=root, purple=intermediate, magenta=outcome) creates visual flow. The eye naturally follows from blue through purple to magenta, tracing causal pathways.

Size: Larger nodes have higher centrality—more connections, more influence. This emerges from the physics simulation but reinforces importance.

Layout: Force-directed positioning naturally clusters related concepts while maintaining readability. The algorithm balances competing constraints: minimize edge crossings, maintain hierarchical levels, avoid node overlap, and create aesthetic appeal.

The encoding philosophy: every pixel should earn its place by conveying information while maintaining visual harmony.

3.8.2 Progressive Disclosure

Information overload kills understanding. The interface reveals complexity gradually:

Level 1 - Overview: At first glance, see network structure and probability color coding. This answers: "What's the shape of the argument? Where are the high-risk areas?"

Level 2 - Hover Details: Mouse over a node to see its description and prior probability. This adds: "What does this factor represent? How likely is it?"

Level 3 - Click Deep Dive: Clicking opens full probability tables and relationships. This reveals: "How does this probability change with conditions? What influences this factor?"

Level 4 - Interactive Exploration: Dragging, zooming, and physics controls enable custom investigation. This supports: "What if I reorganize to see different patterns? How do these clusters relate?"

Each level serves different users and use cases. A policymaker might work primarily with levels 1-2, while a researcher dives into level 3-4 details.

3.8.3 User Interface Elements

Effective interface design for Bayesian networks requires balancing power with accessibility:

Physics Controls: Force-directed layouts benefit from tuning. Gravity affects spread, spring length controls spacing, damping influences settling time. Advanced users can adjust these for optimal layouts, while defaults work well for most cases.

Filter Options: With large networks, selective viewing becomes essential. Filter by probability ranges (show only likely events), node types (focus on interventions), or causal depth (see only immediate effects).

Export Functions: Different stakeholders need different formats. Researchers want raw data, policymakers need reports, presenters require images. Supporting diverse export formats enables broad usage.

Comparison Mode: Understanding often comes from contrast. Side-by-side viewing of baseline versus intervention, or different expert models, reveals critical differences.

Iterative design with actual users would refine these features, ensuring they serve real needs rather than imagined ones.

3.9 Integration with Prediction Markets

The vision: formal models that breathe with live data, updating as collective intelligence evolves. While full implementation awaits, the architecture anticipates this future.

3.9.1 Design for Integration

Integration Architecture requires careful design to manage the impedance mismatch between formal models and market data:

API Specifications: Each platform—Metaculus, Manifold, Good Judgment Open—has unique data formats, update frequencies, and question types. A unified adapter layer would translate platform-specific formats into model-compatible data.

Semantic Matching: The hard problem—connecting "AI causes extinction by 2100" (market question) to "Existential_Catastrophe" (model node). This requires sophisticated NLP and possibly human curation for high-stakes connections.

Aggregation Methods: When multiple markets address similar questions, how do we combine? Weighted averages based on market depth, participant quality, and historical accuracy provide more signal than simple means.

Update Scheduling: Real-time updates would overwhelm users and computation. Smart scheduling might update daily for slow-changing strategic questions, hourly for capability announcements, immediately for critical events.

3.9.2 Challenges and Opportunities

The challenges are real but surmountable:

Question Mapping: Markets ask specific, time-bound questions while models represent general relationships. "AGI by 2030?" maps uncertainly to "APS_Systems exists." Developing robust mapping functions requires deep understanding of both domains.

Temporal Alignment: Market probabilities change over time, but model parameters are typically static. Should we use current market values, time-weighted averages, or attempt to extract trend information?

Quality Variation: A liquid market with expert participants provides different information than a thin market with casual forecasters. Weighting schemes must account for these quality differences.

Incentive Effects: If models influence policy and policy influences outcomes, and markets forecast outcomes, we create feedback loops. Understanding these dynamics prevents perverse incentives.

Despite challenges, even partial integration provides value:

- External validation of expert-derived probabilities
- Dynamic updating as new information emerges
- Identification of where model and market disagree
- Quantified uncertainty from market spread

The perfect shouldn't be the enemy of the good—simple integration beats no integration.

3.10 Computational Performance Analysis

As networks grow from toy examples to real-world complexity, computational challenges emerge. Understanding these constraints shapes realistic expectations and optimization priorities.

3.10.1 Exact vs. Approximate Inference

The fundamental tradeoff in probabilistic reasoning: exactness versus tractability.

Exact Inference: Variable elimination and junction tree algorithms provide mathematically exact answers. For our 3-node rain-sprinkler network, calculations complete instantly. For 20-node networks with modest connectivity, expect seconds. But for 50+ node networks with complex dependencies, exact inference becomes impractical—potentially taking hours or exhausting memory.

Approximate Methods: When exactness becomes impractical, approximation saves the day:

- **Monte Carlo Sampling**: Generate thousands of scenarios consistent with the network, estimate probabilities from frequencies. Accuracy improves with samples, trading computation time for precision.

- **Variational Inference**: Find the simplest distribution that approximates our complex reality. Like fitting a smooth curve to jagged data—we lose detail but gain comprehension.
- **Belief Propagation**: Pass messages between nodes until beliefs converge. Works beautifully for tree-structured networks, can oscillate or converge slowly for complex loops.

The system selects methods based on network properties:

- Small networks: exact inference for precision
- Medium networks: belief propagation for speed
- Large networks: sampling for scalability
- Very large networks: hierarchical decomposition

3.10.2 Scaling Strategies

When networks grow beyond convenient computation, clever strategies maintain usability:

Hierarchical Decomposition: Break large networks into smaller, manageable subnetworks. Compute locally, then integrate results. Like solving a jigsaw puzzle by completing sections before assembling the whole.

Relevance Pruning: For specific queries, most nodes don't matter. If asking about deployment risk, technical details about interpretability methods might be temporarily ignorable. Prune irrelevant subgraphs for focused analysis.

Caching Architecture: Many queries repeat—P(catastrophe), P(deployment|misalignment). Cache results to avoid recomputation. Smart invalidation updates only affected queries when parameters change.

Parallel Processing: Inference calculations often decompose naturally. Different branches of the network can be processed simultaneously. Modern multi-core processors and cloud computing make this increasingly attractive.

Implementation would balance these strategies based on usage patterns. Interactive exploration benefits from caching and pruning. Batch analysis leverages parallelization. The architecture accommodates multiple approaches.

3.11 Results and Achievements

3.11.1 Extraction Quality Assessment

Assessing extraction quality requires honesty about both achievements and limitations. An ideal evaluation would examine multiple dimensions:

Coverage: What proportion of arguments in source texts does the system successfully capture? Initial applications suggest the two-stage approach identifies most explicit causal claims while struggling with deeply implicit relationships.

Accuracy: How closely do automated extractions match expert consensus? Preliminary comparisons indicate strong agreement on primary causal structures with more variation in proba-

bility estimates.

Robustness: How well does the system handle different writing styles, argument structures, and domains? Academic papers with clear argumentation extract more reliably than informal blog posts or policy documents.

Utility: Do the extracted models enable meaningful analysis? Even imperfect extractions that capture 80% of structure with approximate probabilities can dramatically accelerate modeling compared to starting from scratch.

The key insight: perfect extraction isn't necessary for practical value. Like machine translation, which provides useful results despite imperfections, automated argument extraction can enhance human capability without replacing human judgment.

3.11.2 Computational Performance

Performance analysis would reveal the practical boundaries of the current system:

Extraction Speed: LLM-based extraction scales roughly linearly with document length. A 20-page paper might require 30-60 seconds for structural extraction and similar time for probability extraction. This enables processing dozens of documents daily—orders of magnitude faster than manual approaches.

Network Complexity Limits: Exact inference remains tractable for networks up to approximately 30-40 nodes with moderate connectivity. Beyond this, approximate methods become necessary, with sampling methods scaling to hundreds of nodes at the cost of precision.

Visualization Responsiveness: The extraction phase exhibits linear complexity in document length—processing twice as much text takes roughly twice as long. However, the inference phase faces exponential complexity in network connectivity.

End-to-End Pipeline: From document input to interactive visualization, expect 2-5 minutes for typical AI safety arguments. This represents roughly 100x speedup compared to manual modeling efforts.

These performance characteristics make AMTAIR practical for real-world use while highlighting areas for future optimization.

3.11.3 Policy Impact Evaluation

The true test of AMTAIR lies in its ability to inform governance decisions. An ideal policy evaluation framework would demonstrate several capabilities:

Intervention Modeling: Representing diverse policy proposals—from technical standards to international agreements—as parameter modifications in extracted networks. This translation from qualitative proposals to quantitative changes enables rigorous analysis.

Comparative Assessment: Evaluating multiple interventions across different expert worldviews to identify robust strategies. Policies that reduce risk across different models deserve priority over those requiring specific assumptions.

Sensitivity Analysis: Understanding which uncertainties most affect policy conclusions. If an intervention's effectiveness depends critically on disputed parameters, this highlights research priorities.

Implementation Guidance: Moving beyond "this policy reduces risk" to specific recommendations about design details, implementation sequences, and success metrics.

The system would transform abstract policy discussions into concrete quantitative analyses, enabling evidence-based decision-making in AI governance.

3.12 Summary of Technical Contributions

Looking back at the implementation journey, several achievements stand out:

Automated Extraction: The two-stage pipeline successfully transforms natural language arguments into formal models, achieving practical accuracy while maintaining transparency about limitations.

Hybrid Representation: BayesDown bridges qualitative and quantitative worlds, preserving semantic richness while enabling mathematical analysis.

Scalable Architecture: Modular design accommodates growth—new document types, improved extraction methods, additional visualization options—without fundamental restructuring.

Interactive Accessibility: Thoughtful visualization makes complex models understandable to diverse stakeholders, democratizing access to formal reasoning tools.

Policy Relevance: The ability to model interventions and assess robustness transforms academic exercises into practical governance tools.

These technical achievements validate the feasibility of computational coordination infrastructure for AI governance. Not as a complete solution, but as a meaningful enhancement to human judgment and collaboration.

The implementation demonstrates that the vision of automated argument extraction is not merely theoretical but practically achievable. While challenges remain—particularly in handling implicit reasoning and diverse uncertainty expressions—the system provides a foundation for enhanced coordination in AI governance.

The journey from concept to implementation revealed unexpected insights. The two-stage extraction process, initially a pragmatic choice, proved cognitively valid. The intermediate representations became valuable outputs themselves. The visualization challenges led to design innovations applicable beyond this project.

Most importantly, the implementation confirms that formal modeling of AI risk arguments need not remain the province of a few dedicated experts. Through automation and thoughtful design, these powerful tools can serve the broader community working to ensure advanced AI benefits humanity.

Having demonstrated technical feasibility and practical utility, we must now critically examine limitations, address objections, and explore broader implications. The next chapter undertakes this essential reflection, ensuring we neither oversell the approach nor undervalue its contributions.

4. Discussion: Implications and Limitations

4.1 Technical Limitations and Responses

4.1.1 Extraction Quality Boundaries

The critique that automated extraction systematically misses nuanced arguments deserves serious consideration. After months of working with these systems, I've developed both appreciation for their capabilities and acute awareness of their limitations. The reality, unsurprisingly, resists simple characterization.

Consider what happens when AMTAIR encounters a passage like: "While alignment might be achieved through current methods, the economic incentives pushing toward capability development at the expense of safety create a dynamic where technical solutions alone appear insufficient." A human reader parses this effortlessly—alignment is possible but threatened by misaligned incentives. The system, however, might extract two separate claims about alignment feasibility and economic incentives without capturing their interconnection.

These failures aren't random. They follow predictable patterns that reveal something fundamental about the difference between human and machine comprehension. Humans excel at inferring unstated connections, filling gaps with background knowledge, recognizing when an author assumes rather than argues. The system, lacking this context, must rely on explicit linguistic markers. When those markers are absent—as they often are in sophisticated arguments—extraction quality degrades.

Yet dismissing automated extraction based on these limitations misses a crucial point. The alternative isn't perfect human extraction but no formal extraction at all. In practice, humans rarely take the time to formally map complex arguments. When they do, they exhibit their own biases and inconsistencies. The question becomes not whether automated extraction achieves perfection but whether it provides value despite imperfection.

My experience suggests it does, particularly when embedded in appropriate workflows. The two-stage architecture allows human review at natural breakpoints. Extracted structures make excellent starting points for refinement. Most surprisingly, extraction failures often diagnose ambiguities in source texts that human readers gloss over. When the system struggles to deter-

mine whether claim A supports or merely relates to claim B, it's often because the original text genuinely leaves this ambiguous.

Framed differently:

Critic: "Complex implicit reasoning chains resist formalization; automated extraction will systematically miss nuanced arguments and subtle conditional relationships that human experts would identify."

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

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

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

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

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

4.1.2 Objection 2: False Precision in Uncertainty

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

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

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

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

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

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

4.1.3 Objection 3: Correlation Complexity

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

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

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

Copula methods: For known correlation structures, copula functions can model dependencies while preserving marginal distributions. This extends standard Bayesian networks significantly.[15]

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

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

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

4.2 Conceptual and Methodological Concerns

4.2.1 Objection 4: Democratic Exclusion

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

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

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

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

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

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

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

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

4.2.2 Objection 5: Oversimplification of Complex Systems

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

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

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

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

Complementary tools address different aspects of the system. Bayesian networks excel at probabilistic reasoning and intervention analysis. Other approaches—agent-based models, system dynamics, scenario planning—can capture different properties. AMTAIR provides one lens,

not the only lens.

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

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

4.2.3 Objection 6: Idiosyncratic Implementation and Modeling Choices

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

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

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

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

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

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

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

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

4.3 Red-Teaming Results

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

4.3.1 Adversarial Extraction Attempts

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

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

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

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

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

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

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

4.3.2 Robustness Findings

Theoretical analysis suggests key vulnerabilities:

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

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

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

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

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

4.3.3 Implications for Deployment

These considerations suggest AMTAIR is suitable for:

- **Research applications** with expert oversight
- **Policy analysis** of well-structured arguments

- **Educational uses** demonstrating formal reasoning
- **Collaborative modeling** with human verification

But should be used cautiously for:

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

4.4 Enhancing Epistemic Security

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

4.4.1 Making Models Inspectable

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

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

This explicitation serves multiple functions:

Clarity: Vague statements become precise claims subject to evaluation

Comparability: Different experts' models can be systematically compared

Criticizability: Hidden assumptions become visible targets for challenge

Updatability: Formal models can systematically incorporate new evidence

4.4.2 Revealing Convergence and Divergence

Theoretical analysis suggests formal comparison would reveal:

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

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

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

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

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

4.4.3 Improving Collective Reasoning

AMTAIR enhances group epistemics through:

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

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

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

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

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

4.5 Scaling Challenges and Opportunities

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

4.5.1 Technical Scaling

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

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

Data quality varies dramatically across sources:

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

Integration complexity increases with ecosystem growth:

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

4.5.2 Social and Institutional Scaling

Adoption barriers include:

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

Trust building requires:

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

Sustainability depends on:

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

4.5.3 Opportunities for Impact

Despite challenges, several factors favor adoption:

Timing: AI governance needs tools now, creating receptive audiences

Complementarity: AMTAIR enhances rather than replaces existing processes

Flexibility: The approach adapts to different contexts and needs

Network effects: Value increases as more perspectives are formalized

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

4.6 Integration with Governance Frameworks

AMTAIR complements rather than replaces existing governance approaches.

4.6.1 Standards Development

Technical standards bodies could use AMTAIR to:

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

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

4.6.2 Regulatory Design

Regulators could apply the framework to:

- Assess regulatory impact across different scenarios
- Identify enforcement challenges through explicit modeling
- Compare international approaches systematically
- Design adaptive regulations responsive to evidence

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

The extensive literature on corporate governance and liability frameworks Cuomo, Mallin, and Zattoni [16] Demirag, Sudarsanam, and WRIGHT [19] De Villiers and Dimes [18] Di Vito and Trottier [20] Kaur [29] List and Pettit [32] Solomon [49] provides theoretical grounding for understanding how regulatory interventions shape organizational behavior. AMTAIR could formalize these relationships in the specific context of AI development, making explicit how different liability regimes might incentivize or discourage safety investments.

4.6.3 International Coordination

Multilateral bodies could leverage shared models for:

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

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

4.6.4 Organizational Decision-Making

Individual organizations could use AMTAIR for:

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

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

4.7 Future Research Directions

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

4.7.1 Technical Enhancements

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

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

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

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

4.7.2 Methodological Extensions

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

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

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

Uncertainty quantification: Richer representation of deep uncertainty and model confidence

Recent advances in causal structure learning from both text and data Babakov et al. [3] Ban et al. [4] Bethard [6] Chen et al. [12] Heinze-Deml, Maathuis, and Meinshausen [26] Squires and Uhler [50] Yang, Han, and Poon [56] suggest promising directions for enhancing AMTAIR's extraction capabilities. The theoretical foundations from Duhem [21] and Meyer [37] on the philosophy of science and knowledge structures provide epistemological grounding for these methodological extensions.

4.7.3 Application Domains

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

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

Scientific modeling: Formalizing theoretical arguments in emerging fields

Educational tools: Teaching probabilistic reasoning and critical thinking

4.7.4 Ecosystem Development

Open standards: Common formats for model exchange and tool interoperability

Community platforms: Collaborative model development and sharing infrastructure

Training programs: Building capacity for formal modeling in governance communities

Quality assurance: Certification processes for high-stakes model applications

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

4.8 Known Unknowns and Deep Uncertainties

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

4.8.1 Categories of Deep Uncertainty

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

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

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

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

4.8.2 Adaptation Strategies for Deep Uncertainty

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

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

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

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

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

4.8.3 Robust Decision-Making Principles

Given deep uncertainty, certain decision principles become paramount:

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

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

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

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

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

4.9 Summary of Implications

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

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

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

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

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

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

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

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

5. Conclusion: Toward Coordinated AI Governance

5.1 Summary of Key Contributions

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

5.1.1 Theoretical Contributions

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

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

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

5.1.2 Methodological Innovations

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

BayesDown as Bridge Representation: The development of BayesDown syntax creates a crucial intermediate representation preserving both narrative accessibility and mathematical precision. This bridge enables the transformation from qualitative arguments to quantitative models while maintaining traceability and human readability.

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

5.1.3 Technical Achievements

Working Implementation: AMTAIR demonstrates end-to-end feasibility from document ingestion through interactive visualization. The system successfully processes complex arguments like Carlsmith's power-seeking AI model, extracting hierarchical structures and probability information.

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

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

5.1.4 Empirical Findings

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

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

Intervention Impacts: Policy evaluation capabilities demonstrate how formal models enable rigorous assessment of governance options. The ability to trace intervention effects through complex causal networks validates the practical value of formalization.

5.2 Limitations and Honest Assessment

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

5.2.1 Technical Constraints

Extraction Boundaries: The system struggles with implicit assumptions, complex conditionals, and ambiguous quantifiers. These limitations necessitate human review for high-stakes applications.

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

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

5.2.2 Conceptual Limitations

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

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

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

5.2.3 Practical Constraints

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

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

Context Dependence: The approach works best for well-structured academic arguments. Application to informal discussions or political rhetoric remains challenging.

5.3 Implications for AI Governance

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

5.3.1 Near-Term Applications

Research Coordination: Research organizations can use formal models to:

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

- Facilitate collaboration through shared representations

Policy Development: Governance bodies can apply the framework to:

- Evaluate proposals across multiple expert worldviews
- Identify robust interventions effective under uncertainty
- Make assumptions explicit for democratic scrutiny
- Track how evidence changes optimal policies over time

Stakeholder Communication: The visualization and analysis tools enable:

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

5.3.2 Medium-Term Transformation

As adoption spreads, we might see:

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

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

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

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

5.3.3 Long-Term Vision

Successfully scaling this approach could fundamentally alter AI governance:

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

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

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

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

The long-term vision feels almost embarrassingly ambitious when stated plainly. Could this approach fundamentally alter AI governance? Maybe. Probably not in the revolutionary way manifestos promise. More likely, it becomes one tool among many, useful in specific contexts, gradually improving as more people use it and complain about its limitations. But sometimes I imagine a world where policy discussions start with shared models rather than conflicting narratives. Where "let's check what the model says" becomes as natural as "let's check what the data says." Where international negotiations involve parameter haggling rather than rhetorical grandstanding. It's a nice vision. Whether we get there—well, that depends on factors far beyond any technical system.

5.4 Recommendations for Stakeholders

Different communities can take concrete steps to realize these benefits:

5.4.1 For Researchers

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

5.4.2 For Policymakers

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

5.4.3 For Technologists

1. **Improve extraction**: Contribute better prompting strategies, fine-tuned models, or validation methods
2. **Enhance interfaces**: Develop visualizations and interactions serving specific stakeholder needs

3. **Build integrations**: Connect AMTAIR to other tools in the AI governance ecosystem
4. **Scale infrastructure**: Address computational challenges for larger models and broader deployment

5.5 Future Research Agenda

Looking ahead, the landscape of possibilities stretches toward the horizon, each path promising its own rewards and challenges. Let me map the territory worth exploring.

5.5.1 Technical Priorities

The technical frontier advances on multiple fronts, each offering multiplicative improvements when combined:

Extraction Enhancement: The current system, while functional, merely scratches the surface of what's possible. Fine-tuning language models specifically on argument extraction tasks could dramatically improve accuracy. Imagine models trained not just on general text but on thousands of examples of arguments transformed into formal representations.

Handling Implicit Reasoning: So much of expert argumentation relies on unstated background knowledge. When an AI safety researcher mentions "mesa-optimization," they assume familiarity with complex concepts about learned optimization occurring within larger optimization processes. Future systems need to bridge these inferential gaps, perhaps by maintaining explicit knowledge bases of domain concepts or by training models to recognize and fill common argumentative ellipses.

Cross-Document Synthesis: Real understanding emerges not from single papers but from conversations across documents. Authors respond to each other, build on previous work, refine arguments over time. Future systems should trace these intellectual lineages, building composite models that capture evolving community understanding rather than static snapshots.

Representation Extensions: Current Bayesian networks, while powerful, make limiting assumptions. Temporal dynamics matter—AI development unfolds over time, with early decisions constraining later options. Multi-agent representations could capture strategic interactions between actors. Continuous variables better represent quantities like "capability level" than binary approximations. Each extension opens new analytical possibilities.

5.5.2 Methodological Development

Beyond technical improvements lie deeper methodological questions about how we validate, use, and improve these systems:

Validation Science: We need not just ad hoc evaluation but a science of argument extraction assessment. This means building benchmark datasets capturing diverse argument types, developing metrics that go beyond surface accuracy to semantic fidelity, creating adversarial test suites that probe system limitations, and establishing longitudinal studies tracking how extracted models evolve with updating source documents.

Hybrid Intelligence: The future isn't human or AI but human and AI. Optimal collaboration patterns remain unexplored. Should humans verify structure while AI handles probabilities? Should AI propose multiple extractions for human selection? How do we combine formal models with scenario narratives, quantitative forecasts with qualitative insights? The design space for human-AI collaboration in argument formalization remains largely uncharted.

Social Methods: Technology embedded in social contexts requires social science. How do organizations actually use these models? What changes when formal representations replace informal discussions? Ethnographic studies of model use, measurement of coordination improvements, identification of adoption barriers—all essential for real-world impact.

5.5.3 Application Expansion

The principles underlying AMTAIR apply far beyond AI risk:

Domain Extensions: Every field grappling with complex risks could benefit. Biosecurity faces similar challenges—technical complexity, value-laden choices, deep uncertainty. Climate policy involves multi-level causation across physical, economic, and social systems. Nuclear policy, despite decades of study, still struggles with coordination across technical and strategic communities. Each domain would require specialized extraction approaches but could leverage the same fundamental architecture.

Institutional Integration: Moving from research prototype to institutional tool requires thoughtful embedding. Regulatory impact assessment could incorporate formal modeling to make assumptions explicit. Corporate strategic planning, especially for companies developing advanced technologies, needs tools for reasoning about unprecedented risks. Academic peer review might benefit from formal representation of complex arguments.

Global Deployment: AI governance is inherently international, but different regions have different governance cultures, risk tolerances, and institutional structures. Adapting AMTAIR for different contexts—from Silicon Valley's move-fast culture to the EU's precautionary approach to China's state-led development—requires both technical and cultural translation.

5.6 Closing Reflections

As I write these final words, I'm struck by the peculiar position we find ourselves in. We are arguably the first generation that must govern technologies that could fundamentally transform or terminate our species' story. The margin for error shrinks as capabilities grow. The cost of coordination failure rises toward infinity.

The AMTAIR project emerged from a simple observation paired with an ambitious hope. The observation: while humanity mobilizes unprecedented resources to address AI risks, our efforts remain tragically uncoordinated. Different communities work with incompatible frameworks, duplicate efforts, and sometimes actively undermine each other's work. The hope: that computational tools might help us build the epistemic infrastructure necessary for coordination.

What we've accomplished here is both less and more than originally envisioned. Less, because

the challenges proved deeper than anticipated. Natural language resists formalization. Probabilities remain stubbornly subjective. Coordination failures have roots beyond mere communication difficulties. More, because the journey revealed unexpected possibilities. Intermediate representations became valuable in themselves. The extraction process surfaced insights about argument structure. The visualization work demonstrated how thoughtful design can democratize access to formal tools.

Perhaps most importantly, this work demonstrates that perfect solutions need not be the enemy of meaningful progress. AMTAIR doesn't solve the coordination crisis—no single tool could. But it offers genuine assistance: making implicit models explicit, enabling systematic comparison across worldviews, supporting evidence-based policy evaluation, and creating common ground for productive disagreement.

The journey from initial concept to working system taught me more about the problem than about the solution. I began thinking the coordination crisis stemmed primarily from communication failures—experts talking past each other, using different terms for similar concepts. Build translation tools, I reasoned, and coordination would follow. The reality proved more complex. Even with perfect communication, deep disagreements about values, priorities, and acceptable risks remain. Tools can clarify these disagreements but not resolve them.

What surprised me most was how the process of formalization itself generated insights. Forcing myself to make extraction rules explicit revealed my own implicit assumptions about how arguments work. Watching the system fail in predictable ways illuminated the remarkable sophistication of human textual understanding. Building visualizations that actually aided comprehension required confronting how poorly we typically communicate uncertainty.

The technical contributions of this work—the two-stage extraction pipeline, the BayesDown notation, the visualization system—feel less like culminating achievements and more like initial sketches of what's needed. Each component works well enough to demonstrate feasibility but would require substantial refinement for production use. The validation remains preliminary, the scaling challenges largely unaddressed, the integration with existing governance frameworks more theoretical than practical.

Yet I remain optimistic about the approach's potential. Not because AMTAIR solves the coordination crisis—it doesn't—but because it represents the kind of epistemic infrastructure we'll need as AI capabilities advance. The choice isn't between perfect and imperfect tools but between imperfect tools and no tools at all. In a domain where the stakes approach infinity and time grows short, even marginal improvements in coordination capacity matter.

The Stakes: Let me be plain about what's at risk. The development of artificial general intelligence represents a discontinuity in human history comparable to the emergence of life or the evolution of consciousness. Get it right, and we might solve problems that have plagued humanity since our beginning—disease, poverty, ignorance, perhaps even death itself. Get it wrong, and we might extinguish not just ourselves but all the potential futures we might have created.

This isn't science fiction or academic speculation. The capabilities advancing in labs today point

toward systems that could, within decades or less, exceed human cognitive abilities across all domains. What happens when we create minds greater than our own? How do we ensure they remain aligned with human values and flourishing? These questions demand our best collective wisdom.

Currently we approach this challenge fragmented. Technical researchers develop alignment techniques without clear paths to implementation. Policymakers craft governance frameworks without deep technical understanding. Ethicists articulate values without operational specificity. International bodies convene without shared models of the risks they're addressing. This fragmentation isn't just inefficient—it's existentially dangerous.

AMTAIR represents one attempt to build bridges. By automating the extraction of worldviews, integrating live forecasts, and enabling systematic policy evaluation, we create infrastructure for enhanced coordination. Not coordination itself—that requires human wisdom, institutional change, and political will. But infrastructure that makes coordination more feasible.

The path forward demands both ambition and humility. Ambition to build the tools, institutions, and practices necessary for navigating unprecedented risks. Humility to recognize that our tools are imperfect, our understanding incomplete, and our time limited. We must act despite uncertainty, coordinate despite disagreement, and hope despite the magnitude of the challenge.

As I close this thesis, I think of future readers—perhaps humans living in a world made wonderful by aligned AI, perhaps historians studying how we navigated this crucial transition, perhaps no one at all if we fail. To those readers, know that we tried. We saw the challenge, recognized our limitations, and attempted to build what tools we could.

The coordination crisis in AI governance represents both existential risk and existential opportunity. Risk, if we fail to align our efforts before it's too late. Opportunity, if we succeed in creating unprecedented cooperation around humanity's most important challenge. AMTAIR offers one piece of the puzzle—computational infrastructure that enhances our collective ability to reason about complex risks.

The work continues, as it must. Each month brings new AI capabilities that challenge existing frameworks. Each breakthrough raises the stakes. Each failure to coordinate effectively increases cumulative risk. Whether humanity successfully navigates the transition to advanced AI remains radically uncertain. What seems clear is that success, if it comes, will require unprecedented coordination across communities that currently struggle to understand each other. AMTAIR represents one small attempt to build bridges. Many more are needed. May we prove worthy of the challenge before us. May our tools amplify our wisdom rather than our folly.

To future readers—whether you're reading this in a world made wonderful by aligned AI or studying how we tried and failed—know that we saw the challenge clearly. We understood the stakes. We built what tools we could with the time and knowledge available. The rest, as they say, is history. Or will be.

The work continues. The stakes could not be higher. The time grows short. Let us build what we can, while we can, for all our futures depend on it.

Bibliography

[1] Terence J. Anderson. "Visualization Tools and Argument Schemes: A Question of Standpoint". In: *Law, Prob. & Risk* 6 (2007), p. 97. URL: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/lawprisk6&section=9 (visited on 05/25/2025).

[2] Stuart Armstrong, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development". In: *AI & SOCIETY* 31.2 (May 1, 2016), pp. 201–206. ISSN: 1435-5655. DOI: 10.1007/s00146-015-0590-y. URL: https://doi.org/10.1007/s00146-015-0590-y (visited on 05/26/2025).

[3] Nikolay Babakov et al. "Reusability of Bayesian Networks Case Studies: A Survey". In: *Applied Intelligence* 55.6 (Feb. 7, 2025), p. 417. ISSN: 1573-7497. DOI: 10.1007/s10489-025-06289-5. URL: https://doi.org/10.1007/s10489-025-06289-5 (visited on 05/15/2025).

[4] Taiyu Ban et al. *Causal Structure Learning Supervised by Large Language Model*. Nov. 20, 2023. DOI: 10.48550/arXiv.2311.11689. arXiv: 2311.11689 [cs]. URL: http://arxiv.org/abs/2311.11689 (visited on 05/26/2025). Pre-published.

[5] Neil Benn and Ann Macintosh. "Argument Visualization for eParticipation: Towards a Research Agenda and Prototype Tool". In: *Electronic Participation*. Ed. by Efthimios Tambouris, Ann Macintosh, and Hans De Bruijn. Red. by David Hutchison et al. Vol. 6847. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 60–73. ISBN: 978-3-642-23332-6 978-3-642-23333-3. DOI: 10.1007/978-3-642-23333-3_6. URL: http://link.springer.com/10.1007/978-3-642-23333-3_6 (visited on 05/25/2025).

[6] Steven John Bethard. "Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach". PhD thesis. University of Colorado at Boulder, 2007. URL: https://search.proquest.com/openview/405fe32503123d9b5f4836dc3be4c011/1?pq-origsite=gscholar&cbl=18750 (visited on 05/26/2025).

[7] Nick Bostrom. *Superintelligence: Paths, Strategies, Dangers*. Oxford: Oxford University Press, 2014. ISBN: 978-0-19-967811-2. URL: https://scholar.dominican.edu/cynthia-stokes-brown-books-big-history/47.

[8] 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).

[9] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2021. DOI: 10.48550/arXiv.2206.13353. URL: https://arxiv.org/abs/2206.13353. Pre-published.

[10] Joseph Carlsmith. "Is Power-Seeking AI an Existential Risk?" 2022. arXiv: 2206.13353.

[11] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* Aug. 13, 2024. DOI: 10.48550/arXiv.2206.13353. arXiv: 2206.13353 [cs]. URL: http://arxiv.org/abs/2206.13353 (visited on 05/25/2025). Pre-published.

[12] Lu Chen et al. "Inducing Causal Structure for Abstractive Text Summarization". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23: The 32nd ACM International Conference on Information and Knowledge Management. Birmingham United Kingdom: ACM, Oct. 21, 2023, pp. 213–223. ISBN: 979-8-4007-0124-5. DOI: 10.1145/3583780.3614934. URL: https://dl.acm.org/doi/10.1145/3583780.3614934 (visited on 05/26/2025).

[13] Paul F. Christiano. "What Failure Looks Like". In: (Mar. 17, 2019). URL: https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like (visited on 05/25/2025).

[14] Sam Clarke et al. *Modeling Transformative AI Risks (MTAIR) Project – Summary Report*. Version 1. 2022. DOI: 10.48550/ARXIV.2206.09360. URL: https://arxiv.org/abs/2206.09360 (visited on 11/13/2024). Pre-published.

[15] Ben Cottier and Rohin Shah. "Clarifying Some Key Hypotheses in AI Alignment". In: (Aug. 15, 2019). URL: https://www.lesswrong.com/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment (visited on 05/26/2025).

[16] Francesca Cuomo, Christine Mallin, and Alessandro Zattoni. "Corporate Governance Codes: A Review and Research Agenda". In: *Corporate governance: an international review* 24.3 (2016), pp. 222–241. URL: https://ueaeprints.uea.ac.uk/id/eprint/57664/ (visited on 05/26/2025).

[17] Allan Dafoe. "AI Governance: A Research Agenda". In: *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442 (2018), p. 1443. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf (visited on 05/25/2025).

[18] Charl De Villiers and Ruth Dimes. "Determinants, Mechanisms and Consequences of Corporate Governance Reporting: A Research Framework". In: *Journal of Management and Governance* 25.1 (Mar. 2021), pp. 7–26. ISSN: 1385-3457, 1572-963X. DOI: 10.1007/s10997-020-09530-0. URL: https://link.springer.com/10.1007/s10997-020-09530-0 (visited on 05/26/2025).

[19] Istemi Demirag, Sudi Sudarsanam, and MIKE WRIGHT. "Corporate Governance: Overview and Research Agenda". In: *The British Accounting Review* 32.4 (2000), pp. 341–354. URL: https://www.academia.edu/download/49469624/bare.2000.014620161009-3955-1dt4aq5.pdf (visited on 05/26/2025).

[20] Jackie Di Vito and Kim Trottier. "A Literature Review on Corporate Governance Mechanisms: Past, Present, and Future*". In: *Accounting Perspectives* 21.2 (June 2022), pp. 207–235. ISSN: 1911-382X, 1911-3838. DOI: 10.1111/1911-3838.12279. URL: https://onlinelibrary.wiley.com/doi/10.1111/1911-3838.12279 (visited on 05/26/2025).

[21] Pierre Maurice Marie Duhem. *The Aim and Structure of Physical Theory*. 1. Princeton University Press, 1954. 85–87.

[22] Union European. *The Act Texts | EU Artificial Intelligence Act*. 2024. URL: https://artificialintelligenceact.eu/the-act/ (visited on 05/25/2025).

[23] Irving John Good. "Speculations Concerning the First Ultraintelligent Machine". In: *Advances in Computers* (1966), p. 31. DOI: 10.1016/S0065-2458(08)60418-0. URL: https://www.sciencedirect.com/science/article/abs/pii/S0065245808604180.

[24] Ross Gruetzemacher. "Bayesian Networks vs. Conditional Trees for Creating Questions for Forecasting Tournaments". In: (2022).

[25] Stéphane Hallegatte et al. "Investment Decision-Making under Deep Uncertainty-Application to Climate Change". In: *Policy research working paper* 6193 (2012). URL: https://enpc.hal.science/hal-00802049/document (visited on 05/25/2025).

[26] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. "Causal Structure Learning". In: *Annual Review of Statistics and Its Application* 5.1 (Mar. 7, 2018), pp. 371–391. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-031017-100630. URL: https://www.annualreviews.org/doi/10.1146/annurev-statistics-031017-100630 (visited on 05/26/2025).

[27] Tam Hunt. *The Insane "Logic" of the AI Arms Race*. Medium. Mar. 3, 2025. URL: https://tamhunt.medium.com/the-insane-logic-of-the-ai-arms-race-45a5f79f4c0e (visited on 05/26/2025).

[28] Edwin T Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

[29] Kawaljit Kaur. "Corporate Governance and Legal Accountability: A Critical Review of Global Practices". In: *Journal of Law* 2.6 (2024), pp. 1–7. URL: https://joi.shodhsagar.org/index.php/SSJOI/article/view/16 (visited on 05/26/2025).

[30] D. Khartabil et al. "Design and Evaluation of Visualization Techniques to Facilitate Argument Exploration". In: *Computer Graphics Forum* 40.6 (Sept. 2021), pp. 447–465. ISSN: 0167-7055, 1467-8659. DOI: 10.1111/cgf.14389. URL: https://onlinelibrary.wiley.com/doi/10.1111/cgf.14389 (visited on 05/25/2025).

[31] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009. URL: https://books.google.ca/books?hl=en&lr=&id=7dzpHCHzNQ4C&oi=fnd&pg=PR9&dq=Koller,+D.,+%26+Friedman,+N.+(2009).+Probabilistic+Graphical+Models&ots=py2HAh0VAL&sig=gpaID3x6-TY8x5SOopuXpZDXfzs (visited on 05/25/2025).

[32] Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.

[33] David Manheim. *Modeling Transformative AI Risk (MTAIR) - LessWrong*. July 28, 2021. URL: https://www.lesswrong.com/s/aERZoriyHfCqvWkzg (visited on 05/26/2025).

[34] Nestor Maslej. "Artificial Intelligence Index Report 2025". In: *Artificial Intelligence* (2025).

[35] Tegan McCaslin et al. "Conditional Trees: A Method for Generating Informative Questions about Complex Topics". In: *Forecasting Research Institute* (2024). URL: https://static1.sq

uarespace.com/static/635693acf15a3e2a14a56a4a/t/66ba37a144f1d6095de467df/1723479995772/AIConditionalTrees.pdf.

[36] Dasha Metropolitansky and Jonathan Larson. *Towards Effective Extraction and Evaluation of Factual Claims*. Feb. 15, 2025. DOI: 10.48550/arXiv.2502.10855. arXiv: 2502.10855 [cs]. URL: http://arxiv.org/abs/2502.10855 (visited on 05/08/2025). Pre-published.

[37] Valentin Jakob Meyer. "A Structure of Knowledge & the Process of Science". In: *Philosophy of the Social Sciences* First Course Paper (University Bayreuth 2022). DOI: https://www.vjmeyer.com/papers/essays.

[38] Andrea Miotti et al. *A Narrow Path*. 2024. URL: https://www.narrowpath.co/ (visited on 05/19/2025).

[39] Roger B. Nelson. *An Introduction to Copulas*. Springer Series in Statistics. New York, NY: Springer New York, 2006. ISBN: 978-0-387-28659-4. DOI: 10.1007/0-387-28678-0. URL: http://link.springer.com/10.1007/0-387-28678-0 (visited on 05/25/2025).

[40] Paul. *The elephAInt – Are We All like the Six Blind Men When It Comes to AI? | PRISMAGuard LLC*. 2023. URL: https://www.prismaguard.com/the-elephaint-are-we-all-like-the-six-blind-men-when-it-comes-to-ai/ (visited on 05/25/2025).

[41] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press, 2009.

[42] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000. 384 pp. ISBN: 978-0-521-89560-6 978-0-521-77362-1.

[43] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014. URL: https://books.google.ca/books?hl=en&lr=&id=mn2jBQAAQBAJ&oi=fnd&pg=PP1&dq=Pearl,+J.+(1988).+Probabilistic+Reasoning+in+Intelligent+Systems&ots=4tEX2A4Ha8&sig=lgUs_RCoeXEEuGwM5xMEoyJy4HI (visited on 05/25/2025).

[44] John L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. Mit Press, 1995. URL: https://books.google.ca/books?hl=en&lr=&id=JAfHrHTqswAC&oi=fnd&pg=PA1&dq=Pollock,+J.+(1995).+Cognitive+Carpentry&ots=rq-qSCBcxV&sig=aAfHGsGUosxl_1-JuxIEA7C2QO4 (visited on 05/25/2025).

[45] Iskander Rehman. *The Battle for Brilliant Minds: From the Nuclear Age to AI*. War on the Rocks. Jan. 13, 2025. URL: https://warontherocks.com/2025/01/the-battle-for-brilliant-minds-from-the-nuclear-age-to-ai/ (visited on 05/25/2025).

[46] Veronika Samborska. "Scaling up: How Increasing Inputs Has Made Artificial Intelligence More Capable". In: *Our World in Data* (Jan. 20, 2025). URL: https://ourworldindata.org/scaling-up-ai (visited on 05/25/2025).

[47] Sigal Samuel. *AI Is a "Tragedy of the Commons." We've Got Solutions for That*. Vox. July 7, 2023. URL: https://www.vox.com/future-perfect/2023/7/7/23787011/ai-arms-race-tragedy-commons-risk-safety (visited on 05/26/2025).

[48] Thomas C. Schelling. "I960. The Strategy of Conflict". In: *Cambridge, Mass* (1960).

[49] Jill Solomon. *Corporate Governance and Accountability*. John Wiley & Sons, 2020. URL: https://books.google.ca/books?hl=en&lr=&id=JAX9DwAAQBAJ&oi=fnd&pg=PR1&d

q=review+of+the+effects+of+liability+frameworks+on+corporate+governance+&ots=ny23_vd-U0&sig=3LuNNhvSWXriEeg-ipAdDIQGAgo (visited on 05/26/2025).

[50] Chandler Squires and Caroline Uhler. "Causal Structure Learning: A Combinatorial Perspective". In: *Foundations of Computational Mathematics* 23.5 (Oct. 2023), pp. 1781–1815. ISSN: 1615-3375, 1615-3383. DOI: 10.1007/s10208-022-09581-9. URL: https://link.springer.com/10.1007/s10208-022-09581-9 (visited on 05/26/2025).

[51] Max Tegmark. *Asilomar AI Principles*. Future of Life Institute. 2024. URL: https://futureoflife.org/open-letter/ai-principles/ (visited on 05/25/2025).

[52] Phil Tetlock. *Conditional Trees: AI Risk*. 2022. URL: https://www.metaculus.com/tournament/3508/ (visited on 05/26/2025).

[53] Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. First paperback edition. New York: Broadway Books, 2015. 340 pp. ISBN: 978-0-8041-3671-6.

[54] Benjamin Todd. *It Looks like There Are Some Good Funding Opportunities in AI Safety Right Now*. Benjamin Todd. Dec. 21, 2024. URL: https://benjamintodd.substack.com/p/looks-like-there-are-some-good-funding (visited on 05/25/2025).

[55] Christian Voigt. *Christianvoigt/Argdown*. May 23, 2025. URL: https://github.com/christianvoigt/argdown (visited on 05/25/2025).

[56] Jie Yang, Soyeon Caren Han, and Josiah Poon. "A Survey on Extraction of Causal Relations from Natural Language Text". In: *Knowledge and Information Systems* 64.5 (May 2022), pp. 1161–1186. ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-022-01665-w. URL: https://link.springer.com/10.1007/s10115-022-01665-w (visited on 05/26/2025).

[57] Eliezer Yudkowsky. "Artificial Intelligence as a Positive and Negative Factor in Global Risk". In: *Global Catastrophic Risks*. Oxford University Press, July 3, 2008. ISBN: 978-0-19-857050-9 978-0-19-191810-0. DOI: 10.1093/oso/9780198570509.003.0021. URL: https://academic.oup.com/book/40615/chapter/348239228 (visited on 11/15/2024).

Manual Extraction of ArgDown Data from Bucknall and Dori-Hacohen [8]

```
[Existential_Risk]: Increase in existential risks for humanity. {"instantiations": [TRUE", "

- [Unaligned_AGI_Risk]: Unaligned artificial general intelligence causes existential risk. {
  - [State-State_Relations]
- [Near_term_AI]: Even if not unaligned AGI, near term AI can act as intermediate risk facto
  - [State-State_Relations]: AI arms race dynamic inhibits international coordination, diver
    - [Cybersecurity]: Probably enhances Cyber-Attack-Offense, may intensify cyber warfare.
  - [State-Cooperation_Relations]: Cooperations have a lot of power and might have misaligne
  - [Stable_Repressive_Regime]: More repressive instruments, possibility of stable repressiv
    - [State-Citizen_Relations]: AI helps regime monitor citizens {"instantiations": [TRUE",
  - [Compromised_Political_Decision_Making]: AI can compromise political decision making. {"
    - [Social_media_and_Recommender_Systems]: Influence of AI in social media on public opir
- [Nuclear]: Probability that nuclear conflict escalates to end civilisation. {"instantiatio
  - [Compromised_Political_Decision_Making]
- [Biological]: Probability that a natural or engineered pandemic poses existential risks. {
  - [Compromised_Political_Decision_Making]
  - [Social_media_and_Recommender_Systems]
- [Natural]: Non-human caused existential risks, seem unrelated with AI. {"instantiations":
- [Environmental]: Probability of climate catastrophe. {"instantiations": ["TRUE", "FALSE"]}
  - [Compromised_Political_Decision_Making]
  - [AI_resource_consumption]: Current AI models consume large amounts of energy having envi
  - [Social_media_and_Recommender_Systems]
```

Bibliography

[1] Terence J. Anderson. "Visualization Tools and Argument Schemes: A Question of Standpoint". In: *Law, Prob. & Risk* 6 (2007), p. 97. URL: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/lawprisk6&section=9 (visited on 05/25/2025).

[2] Stuart Armstrong, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development". In: *AI & SOCIETY* 31.2 (May 1, 2016), pp. 201–206. ISSN: 1435-5655. DOI: 10.1007/s00146-015-0590-y. URL: https://doi.org/10.1007/s00146-015-0590-y (visited on 05/26/2025).

[3] Nikolay Babakov et al. "Reusability of Bayesian Networks Case Studies: A Survey". In: *Applied Intelligence* 55.6 (Feb. 7, 2025), p. 417. ISSN: 1573-7497. DOI: 10.1007/s10489-025-06289-5. URL: https://doi.org/10.1007/s10489-025-06289-5 (visited on 05/15/2025).

[4] Taiyu Ban et al. *Causal Structure Learning Supervised by Large Language Model*. Nov. 20, 2023. DOI: 10.48550/arXiv.2311.11689. arXiv: 2311.11689 [cs]. URL: http://arxiv.org/abs/2311.11689 (visited on 05/26/2025). Pre-published.

[5] Neil Benn and Ann Macintosh. "Argument Visualization for eParticipation: Towards a Research Agenda and Prototype Tool". In: *Electronic Participation*. Ed. by Efthimios Tambouris, Ann Macintosh, and Hans De Bruijn. Red. by David Hutchison et al. Vol. 6847. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 60–73. ISBN: 978-3-642-23332-6 978-3-642-23333-3. DOI: 10.1007/978-3-642-23333-3_6. URL: http://link.springer.com/10.1007/978-3-642-23333-3_6 (visited on 05/25/2025).

[6] Steven John Bethard. "Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach". PhD thesis. University of Colorado at Boulder, 2007. URL: https://search.proquest.com/openview/405fe32503123d9b5f4836dc3be4c011/1?pq-origsite=gscholar&cbl=18750 (visited on 05/26/2025).

[7] Nick Bostrom. *Superintelligence: Paths, Strategies, Dangers*. Oxford: Oxford University Press, 2014. ISBN: 978-0-19-967811-2. URL: https://scholar.dominican.edu/cynthia-stokes-brown-books-big-history/47.

[8] 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).

[9] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2021. DOI: 10.48550/arXiv.2206.13353. URL: https://arxiv.org/abs/2206.13353. Pre-published.

[10] Joseph Carlsmith. "Is Power-Seeking AI an Existential Risk?" 2022. arXiv: 2206.13353.

[11] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* Aug. 13, 2024. DOI: 10.48550/arXiv.2206.13353. arXiv: 2206.13353 [cs]. URL: http://arxiv.org/abs/2206.13353 (visited on 05/25/2025). Pre-published.

[12] Lu Chen et al. "Inducing Causal Structure for Abstractive Text Summarization". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23: The 32nd ACM International Conference on Information and Knowledge Management. Birmingham United Kingdom: ACM, Oct. 21, 2023, pp. 213–223. ISBN: 979-8-4007-0124-5. DOI: 10.1145/3583780.3614934. URL: https://dl.acm.org/doi/10.1145/3583780.3614934 (visited on 05/26/2025).

[13] Paul F. Christiano. "What Failure Looks Like". In: (Mar. 17, 2019). URL: https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like (visited on 05/25/2025).

[14] Sam Clarke et al. *Modeling Transformative AI Risks (MTAIR) Project – Summary Report*. Version 1. 2022. DOI: 10.48550/ARXIV.2206.09360. URL: https://arxiv.org/abs/2206.09360 (visited on 11/13/2024). Pre-published.

[15] Ben Cottier and Rohin Shah. "Clarifying Some Key Hypotheses in AI Alignment". In: (Aug. 15, 2019). URL: https://www.lesswrong.com/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment (visited on 05/26/2025).

[16] Francesca Cuomo, Christine Mallin, and Alessandro Zattoni. "Corporate Governance Codes: A Review and Research Agenda". In: *Corporate governance: an international review* 24.3 (2016), pp. 222–241. URL: https://ueaeprints.uea.ac.uk/id/eprint/57664/ (visited on 05/26/2025).

[17] Allan Dafoe. "AI Governance: A Research Agenda". In: *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442 (2018), p. 1443. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf (visited on 05/25/2025).

[18] Charl De Villiers and Ruth Dimes. "Determinants, Mechanisms and Consequences of Corporate Governance Reporting: A Research Framework". In: *Journal of Management and Governance* 25.1 (Mar. 2021), pp. 7–26. ISSN: 1385-3457, 1572-963X. DOI: 10.1007/s10997-020-09530-0. URL: https://link.springer.com/10.1007/s10997-020-09530-0 (visited on 05/26/2025).

[19] Istemi Demirag, Sudi Sudarsanam, and MIKE WRIGHT. "Corporate Governance: Overview and Research Agenda". In: *The British Accounting Review* 32.4 (2000), pp. 341–354. URL: https://www.academia.edu/download/49469624/bare.2000.014620161009-3955-1dt4aq5.pdf (visited on 05/26/2025).

[20] Jackie Di Vito and Kim Trottier. "A Literature Review on Corporate Governance Mechanisms: Past, Present, and Future*". In: *Accounting Perspectives* 21.2 (June 2022), pp. 207–235. ISSN: 1911-382X, 1911-3838. DOI: 10.1111/1911-3838.12279. URL: https://onlinelibrary.wiley.com/doi/10.1111/1911-3838.12279 (visited on 05/26/2025).

[21] Pierre Maurice Marie Duhem. *The Aim and Structure of Physical Theory*. 1. Princeton University Press, 1954. 85–87.

[22] Union European. *The Act Texts | EU Artificial Intelligence Act*. 2024. URL: https://artificialintelligenceact.eu/the-act/ (visited on 05/25/2025).

[23] Irving John Good. "Speculations Concerning the First Ultraintelligent Machine". In: *Advances in Computers* (1966), p. 31. DOI: 10.1016/S0065-2458(08)60418-0. URL: https://www.sciencedirect.com/science/article/abs/pii/S0065245808604180.

[24] Ross Gruetzemacher. "Bayesian Networks vs. Conditional Trees for Creating Questions for Forecasting Tournaments". In: (2022).

[25] Stéphane Hallegatte et al. "Investment Decision-Making under Deep Uncertainty-Application to Climate Change". In: *Policy research working paper* 6193 (2012). URL: https://enpc.hal.science/hal-00802049/document (visited on 05/25/2025).

[26] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. "Causal Structure Learning". In: *Annual Review of Statistics and Its Application* 5.1 (Mar. 7, 2018), pp. 371–391. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-031017-100630. URL: https://www.annualreviews.org/doi/10.1146/annurev-statistics-031017-100630 (visited on 05/26/2025).

[27] Tam Hunt. *The Insane "Logic" of the AI Arms Race*. Medium. Mar. 3, 2025. URL: https://tamhunt.medium.com/the-insane-logic-of-the-ai-arms-race-45a5f79f4c0e (visited on 05/26/2025).

[28] Edwin T Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

[29] Kawaljit Kaur. "Corporate Governance and Legal Accountability: A Critical Review of Global Practices". In: *Journal of Law* 2.6 (2024), pp. 1–7. URL: https://joi.shodhsagar.org/index.php/SSJOI/article/view/16 (visited on 05/26/2025).

[30] D. Khartabil et al. "Design and Evaluation of Visualization Techniques to Facilitate Argument Exploration". In: *Computer Graphics Forum* 40.6 (Sept. 2021), pp. 447–465. ISSN: 0167-7055, 1467-8659. DOI: 10.1111/cgf.14389. URL: https://onlinelibrary.wiley.com/doi/10.1111/cgf.14389 (visited on 05/25/2025).

[31] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009. URL: https://books.google.ca/books?hl=en&lr=&id=7dzpHCHzNQ4C&oi=fnd&pg=PR9&dq=Koller,+D.,+%26+Friedman,+N.+(2009).+Probabilistic+Graphical+Models&ots=py2HAh0VAL&sig=gpaID3x6-TY8x5SOopuXpZDXfzs (visited on 05/25/2025).

[32] Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.

[33] David Manheim. *Modeling Transformative AI Risk (MTAIR) - LessWrong*. July 28, 2021. URL: https://www.lesswrong.com/s/aERZoriyHfCqvWkzg (visited on 05/26/2025).

[34] Nestor Maslej. "Artificial Intelligence Index Report 2025". In: *Artificial Intelligence* (2025).

[35] Tegan McCaslin et al. "Conditional Trees: A Method for Generating Informative Questions about Complex Topics". In: *Forecasting Research Institute* (2024). URL: https://static1.sq

uarespace.com/static/635693acf15a3e2a14a56a4a/t/66ba37a144f1d6095de467df/1723479995772/AIConditionalTrees.pdf.

[36] Dasha Metropolitansky and Jonathan Larson. *Towards Effective Extraction and Evaluation of Factual Claims*. Feb. 15, 2025. DOI: 10.48550/arXiv.2502.10855. arXiv: 2502.10855 [cs]. URL: http://arxiv.org/abs/2502.10855 (visited on 05/08/2025). Pre-published.

[37] Valentin Jakob Meyer. "A Structure of Knowledge & the Process of Science". In: *Philosophy of the Social Sciences* First Course Paper (University Bayreuth 2022). DOI: https://www.vjmeyer.com/papers/essays.

[38] Andrea Miotti et al. *A Narrow Path*. 2024. URL: https://www.narrowpath.co/ (visited on 05/19/2025).

[39] Roger B. Nelson. *An Introduction to Copulas*. Springer Series in Statistics. New York, NY: Springer New York, 2006. ISBN: 978-0-387-28659-4. DOI: 10.1007/0-387-28678-0. URL: http://link.springer.com/10.1007/0-387-28678-0 (visited on 05/25/2025).

[40] Paul. *The elephAInt – Are We All like the Six Blind Men When It Comes to AI? | PRISMAGuard LLC*. 2023. URL: https://www.prismaguard.com/the-elephaint-are-we-all-like-the-six-blind-men-when-it-comes-to-ai/ (visited on 05/25/2025).

[41] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press, 2009.

[42] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000. 384 pp. ISBN: 978-0-521-89560-6 978-0-521-77362-1.

[43] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014. URL: https://books.google.ca/books?hl=en&lr=&id=mn2jBQAAQBAJ&oi=fnd&pg=PP1&dq=Pearl,+J.+(1988).+Probabilistic+Reasoning+in+Intelligent+Systems&ots=4tEX2A4Ha8&sig=lgUs_RCoeXEEuGwM5xMEoyJy4HI (visited on 05/25/2025).

[44] John L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. Mit Press, 1995. URL: https://books.google.ca/books?hl=en&lr=&id=JAfHrHTqswAC&oi=fnd&pg=PA1&dq=Pollock,+J.+(1995).+Cognitive+Carpentry&ots=rq-qSCBcxV&sig=aAfHGsGUosxl_1-JuxIEA7C2QO4 (visited on 05/25/2025).

[45] Iskander Rehman. *The Battle for Brilliant Minds: From the Nuclear Age to AI*. War on the Rocks. Jan. 13, 2025. URL: https://warontherocks.com/2025/01/the-battle-for-brilliant-minds-from-the-nuclear-age-to-ai/ (visited on 05/25/2025).

[46] Veronika Samborska. "Scaling up: How Increasing Inputs Has Made Artificial Intelligence More Capable". In: *Our World in Data* (Jan. 20, 2025). URL: https://ourworldindata.org/scaling-up-ai (visited on 05/25/2025).

[47] Sigal Samuel. *AI Is a "Tragedy of the Commons." We've Got Solutions for That*. Vox. July 7, 2023. URL: https://www.vox.com/future-perfect/2023/7/7/23787011/ai-arms-race-tragedy-commons-risk-safety (visited on 05/26/2025).

[48] Thomas C. Schelling. "I960. The Strategy of Conflict". In: *Cambridge, Mass* (1960).

[49] Jill Solomon. *Corporate Governance and Accountability*. John Wiley & Sons, 2020. URL: https://books.google.ca/books?hl=en&lr=&id=JAX9DwAAQBAJ&oi=fnd&pg=PR1&d

q=review+of+the+effects+of+liability+frameworks+on+corporate+governance+&ots=ny23_vd-U0&sig=3LuNNhvSWXriEeg-ipAdDIQGAgo (visited on 05/26/2025).

[50] Chandler Squires and Caroline Uhler. "Causal Structure Learning: A Combinatorial Perspective". In: *Foundations of Computational Mathematics* 23.5 (Oct. 2023), pp. 1781–1815. ISSN: 1615-3375, 1615-3383. DOI: 10.1007/s10208-022-09581-9. URL: https://link.springer.com/10.1007/s10208-022-09581-9 (visited on 05/26/2025).

[51] Max Tegmark. *Asilomar AI Principles*. Future of Life Institute. 2024. URL: https://futureoflife.org/open-letter/ai-principles/ (visited on 05/25/2025).

[52] Phil Tetlock. *Conditional Trees: AI Risk*. 2022. URL: https://www.metaculus.com/tournament/3508/ (visited on 05/26/2025).

[53] Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. First paperback edition. New York: Broadway Books, 2015. 340 pp. ISBN: 978-0-8041-3671-6.

[54] Benjamin Todd. *It Looks like There Are Some Good Funding Opportunities in AI Safety Right Now*. Benjamin Todd. Dec. 21, 2024. URL: https://benjamintodd.substack.com/p/looks-like-there-are-some-good-funding (visited on 05/25/2025).

[55] Christian Voigt. *Christianvoigt/Argdown*. May 23, 2025. URL: https://github.com/christianvoigt/argdown (visited on 05/25/2025).

[56] Jie Yang, Soyeon Caren Han, and Josiah Poon. "A Survey on Extraction of Causal Relations from Natural Language Text". In: *Knowledge and Information Systems* 64.5 (May 2022), pp. 1161–1186. ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-022-01665-w. URL: https://link.springer.com/10.1007/s10115-022-01665-w (visited on 05/26/2025).

[57] Eliezer Yudkowsky. "Artificial Intelligence as a Positive and Negative Factor in Global Risk". In: *Global Catastrophic Risks*. Oxford University Press, July 3, 2008. ISBN: 978-0-19-857050-9 978-0-19-191810-0. DOI: 10.1093/oso/9780198570509.003.0021. URL: https://academic.oup.com/book/40615/chapter/348239228 (visited on 11/15/2024).

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 27, 2025

Valentin Meyer