PY_AMTAIRResearchProposal

Research Proposal: AMTAIR - Automating Transformative AI Risk Modeling

Executive Summary:

The Automating Transformative AI Risk Modeling (AMTAIR) project addresses a critical coordination failure in AI governance: despite unprecedented investment in AI safety, we lack the strategic infrastructure needed to align disparate efforts across technical, governance, and policy domains.

We're developing computational tools that automate the extraction of probabilistic world models from AI safety literature using frontier language models. These tools will form the foundation for a comprehensive, adaptive AI Grand Strategy that remains robust across various futures.

By integrating Bayesian networks, live forecasting data, and automated extraction pipelines, our approach will:

1. Quantify existential risks from AI systems and identify intervention points
2. Make implicit models explicit, highlighting agreements and cruxes of disagreement
3. Evaluate policy impacts across multiple worldviews and scenarios
4. Generate strategic recommendations that adapt as new information emerges

We require approximately $200,000 to build these tools over 12 months, validate them with experts, and develop an initial AI Grand Strategy framework. Our team combines expertise in Bayesian modeling, AI governance, and forecasting theory—bridging technical, social, and policy domains with precisely the interdisciplinary skills needed for this challenge.

The window for establishing effective governance is narrowing as AI capabilities accelerate. Our project creates the epistemic infrastructure necessary for global coordination on what may be humanity's most consequential technical challenge.

1. Introduction and Identification of the Problem:

1.1 The Coordination Problem

The development of advanced AI systems presents a paradox: unprecedented investment in safety research coexists alongside a fundamental coordination failure. Despite millions in funding, rapidly growing awareness, and proliferating frameworks, we lack the strategic infrastructure needed to align these disparate efforts as AI capabilities advance at an accelerating pace.

This isn't merely inefficient—it systematically increases existential risk. When organizations function as independent processors without shared protocols, we generate duplicative work, leave critical gaps unaddressed, and create inconsistent approaches to interdependent problems. Technical

alignment researchers develop solutions without implementation pathways; policy specialists craft frameworks without technical grounding; ethicists articulate principles without operational specificity. Each community operates with different terminologies, priorities, and implicit theories of change—a fragmentation that becomes exponentially more dangerous as capabilities approach human-level intelligence.

Technical solutions alone cannot address these coordination challenges. Perfect alignment techniques would still fail in a misaligned governance landscape where competitive pressures, verification challenges, and international tensions remain unresolved. We need tools that bridge technical and governance domains, making implicit models explicit and facilitating coordination around shared understanding. As we will see next, the consequences of failing to do so compound exponentially as AI capabilities advance.

1.2 The Exponential Risk Multiplier

As AI capabilities advance, the consequences of coordination failures compound rather than merely accumulate. This follows directly from probability theory: when multiple systems must function correctly to avoid catastrophe, the probability of failure increases exponentially with the number of potential failure points.

Consider a simplified model where safety depends on coordinated responses across N domains (technical alignment, deployment oversight, international governance, etc.). If each domain has an independent 10% chance of failure (though in reality, failures are likely correlated), the probability of at least one critical failure scales as:

P(failure) = 1 - (0.9)^N

With just 3 domains this results in: P(failure) = 27.1% and with 5 domains: P(failure) = 41.0%

This exponential scaling creates a rapidly narrowing window for effective intervention. Given current AI development trajectories—where capabilities previously projected decades away emerge within months—we face unprecedented time pressure. The recent compression of milestone achievement intervals (e.g., time between GPT-3 and GPT-4 compared to previous advancements) suggests this window continues to shrink.

Unlike other global coordination problems (climate change, nuclear security), AI development presents unique challenges:

- Threshold dynamics where capabilities may rapidly cross critical thresholds
- Intrinsic advantages to offense over defense in deployment
- Verification challenges that complicate trust mechanisms
- Potential for irreversible deployment decisions

These factors make the coordination problem both more urgent and more difficult than comparable challenges in human history.

1.3 The Information Processing Bottleneck

At its core, AI governance is an information processing challenge. Developing effective strategy requires integrating insights across technical, social, economic, and political domains—each with its own language, assumptions, and uncertainty profiles.

Currently, this integration happens primarily through manual processes:

- Researchers reading and interpreting papers from adjacent fields
- Workshops and conferences facilitating cross-domain exchange
- Individuals mentally synthesizing perspectives into personal world models

These manual approaches create severe bottlenecks as the volume and complexity of relevant information grows. For example, the Modeling Transformative AI Risks (MTAIR) project demonstrated the value of formalizing world models using Bayesian networks, but required intensive labor to extract these models from research papers and expert judgments.

The bottleneck manifests in concrete operational limitations:

- New research takes months to incorporate into strategic thinking
- Expert time is consumed by basic information processing rather than novel insights
- Cross-domain translation relies on rare individuals with interdisciplinary backgrounds
- Models remain static rather than updating with new developments

Crucially, this information processing challenge doesn't scale linearly with more researchers—the complexity of coordination itself grows superlinearly with team size. More researchers without better coordination tools can actually decrease overall effectiveness.

Our project addresses precisely this bottleneck, automating critical components of the information processing pipeline while preserving human judgment where it remains essential. By leveraging frontier LLMs for structured knowledge extraction and integrating with live forecasting platforms, we create a dynamic strategic coordination platform that scales with the growing complexity of the AI governance landscape. This leads us to our central research question: how can we effectively automate and formalize world models from AI governance literature to enable robust prediction of policy impacts?

2. Research Question and Objectives

2.1 Primary Research Question

How can frontier AI technologies be leveraged to automate the extraction and formalization of causal world models from AI governance literature, enabling robust prediction of policy impacts across divergent futures and facilitating improved strategic coordination?

This question directly addresses the coordination problem and information processing bottleneck identified above. By automating the extraction of implicit models from research papers and

formalizing them in a common representational framework, we can overcome key barriers to strategic coordination in AI governance.

The question encompasses several critical dimensions:

- Technical feasibility (can LLMs extract structured causal models with sufficient accuracy?)
- Integration challenges (how can qualitative arguments be translated into quantitative frameworks?)
- Evaluation methods (how do we assess whether extracted models capture the original intent?)
- Application pathways (how can these formalized models inform policy decisions?)

By framing our inquiry in terms of automation, extraction, and formalization, we focus on the specific technical challenges that must be overcome to scale AI governance capacity. The emphasis on "robust prediction" highlights our goal of developing tools that remain valuable across different potential futures rather than optimizing for a single scenario.

2.2 Sub-Questions

To answer our primary research question, we must address several interconnected sub-questions:

1. **How can LLMs effectively extract causal world models from unstructured text?**
    - What prompting strategies maximize extraction accuracy?
    - How can we detect and correct for extraction errors or biases?
    - What intermediate representations facilitate the transition from text to formal models?
    - How do we handle disagreements or contradictions within source materials?
2. **What mechanisms enable reliable integration between prediction markets and Bayesian networks?**
    - How should forecasts be mapped to model variables?
    - What weighting schemes balance different information sources?
    - How frequently should updates occur, and how should conflicting signals be resolved?
    - What feedback loops maintain calibration between models and real-world developments?
3. **How can policy impacts be rigorously modeled across diverse worldviews?**
    - What methodologies identify robust interventions despite worldview disagreements?
    - How should policy interventions be represented within causal models?
    - What metrics effectively capture policy effectiveness across different value systems?
    - How can counterfactual reasoning about policy impacts be validated?
4. **What specific attributes must our modeling tools possess to meaningfully represent existential risk?**
    - How should uncertainty be represented in both model structure and parameters?

    - What level of granularity balances comprehensiveness with usability?
    - How can subjective judgments be integrated with empirical data?
    - What validation methods can assess model quality despite the absence of ground truth?

These sub-questions define concrete research pathways that collectively address our primary question. Each represents a distinct technical challenge that must be solved to realize our vision of automated, formalized world models for AI governance, and together they form the foundation of our methodological approach.

2.3 Success Metrics

We will evaluate the success of our research through multiple measurable criteria:

Extraction Accuracy (Benchmark against Human Expert Annotation)

- Precision and recall of node identification compared with expert annotation
- Precision and recall of edge identification compared with expert annotation
- Probability estimation compared to expert judgments
- Identification of critical variables as judged by domain experts

Model Validation Criteria

- Formal verification of model mathematical consistency (e.g., probabilities sum to 1)
- Structural validity assessment through expert review
- Calibration assessment through historical backtesting where applicable
- Sensitivity analysis to identify and address unstable model components

Usability Metrics for Policy Stakeholders

- Time to understand key model insights for informed non-specialists
- Success rate for common tasks for target user groups
- System Usability Scale (SUS) score for primary interfaces
- Reported intent to adopt from surveyed stakeholders

Posterior Probability Updates

- Formalized likelihood updates in response to new information
- Convergence assessment across multiple worldviews as evidence accumulates
- Surprise minimization measured through proper scoring rules
- Response time to incorporate new research findings

These metrics provide concrete benchmarks to assess our progress and the ultimate utility of our tools. We recognize that some metrics involve subjective judgments (e.g., expert approval ratings), but we will use structured evaluation protocols to ensure consistency and transparency in these assessments.

By setting specific, measurable criteria across technical accuracy, validation, usability, and adaptability dimensions, we create accountability for our research and clear milestones to guide our development process.

3. Theoretical Foundations and State of the Art

3.1 The MTAIR Framework: Achievements and Limitations

The Modeling Transformative AI Risks (MTAIR) project represents the most comprehensive attempt to date at formalizing existential risks from advanced AI. Developed by a team including Clarke, Cottier, Englander, Eth, Manheim, Martin, and Rice, MTAIR created a structured Bayesian network model of key risk factors, uncertainties, and potential interventions.

MTAIR's core achievements include:

- Mapping complex causal relationships between AI development, governance, and risk factors
- Quantifying uncertainty through probabilistic representations
    - Enabling sensitivity analysis to identify crucial variables
- Creating a shared vocabulary for discussing AI risk pathways
- Demonstrating the feasibility of formal modeling for existential risks

However, MTAIR also revealed significant limitations that our project aims to address:

Manual Bottlenecks The original MTAIR implementation required intensive manual labor to extract and formalize world models from research papers and expert judgments. This process involved reading papers, identifying key claims, mapping causal relationships, and estimating probabilities—all completed by human researchers. This approach simply doesn't scale with the growing volume of AI safety literature.

Static Nature Once constructed, updating the MTAIR model with new research findings required substantial manual effort. The model remained largely static rather than dynamically incorporating emerging insights and data. In a rapidly evolving field, this severely limits ongoing relevance.

Limited Accessibility The Analytica software implementation of MTAIR, while powerful, presented barriers to widespread engagement. Complex model structures and interfaces restricted meaningful interaction to those with specialized training, limiting broader uptake among policymakers and other stakeholders.

Worldview Integration Challenges While MTAIR acknowledged diverse perspectives on AI risk, fully representing multiple worldviews proved challenging. The model primarily reflected a synthesis rather than enabling exploration of how different assumptions lead to divergent conclusions—a crucial capability for identifying robust interventions.

Our project directly addresses these limitations by:

1. Automating the extraction process using frontier LLMs

2. Creating dynamic update mechanisms integrated with live data sources
3. Developing intuitive interfaces for different stakeholder groups
4. Explicitly modeling multiple worldviews and their implications

By building on MTAIR's achievements while overcoming its limitations, we position our work as the natural next step in the evolution of AI risk modeling.

3.2 Probabilistic Modeling for AI Safety

Bayesian networks provide the mathematical foundation for our approach to AI risk modeling. These probabilistic graphical models represent variables as nodes and causal relationships as directed edges, with conditional probability tables encoding the strength and nature of these relationships.

Bayesian Networks as Knowledge Representation Bayesian networks offer several advantages for representing knowledge about AI risks:

- They naturally encode uncertainty through probability distributions
- They capture conditional independence relationships, simplifying complex systems
- They support both causal reasoning (effects of interventions) and evidential reasoning (updating based on observations)
- They provide a formal framework for integrating diverse information sources

For AI safety specifically, Bayesian networks enable:

- Decomposition of complex risk scenarios into more tractable components
- Explicit representation of causal pathways from development decisions to outcomes
- Principled updating as new evidence emerges
- Identification of critical uncertainties through sensitivity analysis

Directed Acyclic Graphs (DAGs) and Causal Inference DAGs form the structural backbone of Bayesian networks, with important theoretical properties:

- Acyclicity ensures coherent probability calculations
- Directed edges represent causal relationships rather than mere correlations
- The d-separation criterion identifies conditional independence relationships
- Pearl's do-calculus enables reasoning about interventions

These properties are crucial for modeling AI risk, where we need to distinguish:

- Correlations that merely predict outcomes
- Causal relationships that can be leveraged for intervention
- Confounding variables that might mislead analysis

Current Limitations in Automated Causal Extraction Despite significant advances in natural language processing, automatically extracting causal models from text remains challenging:

- Language often expresses causality implicitly or ambiguously (e.g., "As AI systems become more capable, governance challenges will increase" leaves unclear whether capability directly causes governance challenges or operates through intermediate mechanisms)
- Arguments rely on unstated assumptions and background knowledge
- Probability judgments may be expressed qualitatively rather than quantitatively
- Technical terminology varies across disciplines

Recent work on causal extraction using LLMs has shown promise but still requires careful prompting and validation. Our approach builds on this work while developing specialized techniques for the AI safety domain, where arguments often involve complex counterfactuals and nested conditional statements.

3.3 AI Governance Literature and Forecasting

The AI governance landscape spans multiple disciplinary traditions, each with its own approaches to uncertainty, evidence, and intervention design.

Key AI Governance Frameworks Several frameworks currently shape discourse on AI governance:

- The governance of AI safety principles (e.g., Asilomar principles)
- Risk-based regulatory approaches (e.g., EU AI Act, now adopted but being implemented)
- International coordination mechanisms (e.g., OECD AI principles)
- Technical standards frameworks (e.g., IEEE Ethics frameworks)
- Corporate governance structures (e.g., responsible AI teams)

These frameworks often operate with implicit world models—assumptions about how technological development proceeds, how incentives shape behavior, and how interventions affect outcomes. Making these models explicit is essential for meaningful comparison and integration.

Prediction Markets and Expert Forecasting Forecasting platforms represent a crucial complementary approach to AI governance:

- Metaculus hosts specialized questions on AI development trajectories and risk factors (a primary integration target for our project)
- Good Judgment Project and Samotsvety provide expert forecasts on key questions
- Manifold Markets enables community prediction on more granular questions
- Epochs' AI forecasting tracks real-world progress against predictions

These platforms generate valuable probability estimates for specific questions, but integration with comprehensive causal models remains limited. Current approaches typically:

- Treat forecasts as isolated data points rather than components of a causal system
- Lack formal methods for updating complex models based on forecast results
- Miss opportunities to identify which forecasts would be most informative for decision-making

Gaps Between Technical Risk Modeling and Policy Implementation A critical gap exists between technical risk models and actionable policy:

- Technical models often lack the institutional context necessary for implementation
- Policy frameworks typically underspecify the causal mechanisms by which they reduce risk
- Translation between technical and policy languages remains largely manual
- Feedback loops for policy evaluation based on risk models are underdeveloped

Our project aims to bridge this gap by creating tools that connect technical risk assessments to concrete policy levers, enabling policymakers to explore intervention impacts and technical researchers to frame their work in terms of governance implications.

3.4 Transformative Potential of LLM-Assisted Modeling

Recent advances in large language models have created a technological inflection point that enables our approach for the first time.

Frontier LLM Capabilities for Structured Knowledge Extraction Models like GPT-4, GPT-4o, R1, Claude 3.7 Sonnet, and Claude Opus demonstrate unprecedented capabilities for transforming unstructured text into structured representations:

- They can identify entities, relationships, and claims from complex technical documents
- They understand nuanced expressions of uncertainty and conditionality
- They can maintain consistency across long contexts necessary for modeling complex arguments
- They can translate between different disciplinary languages and frameworks

Our preliminary experiments using carefully engineered prompts show that these models can extract causal structures from AI safety papers with accuracy approaching human performance on many dimensions. For example, when tasked with identifying key variables and causal relationships from Carlsmith's work on power-seeking AI, Claude 3.7 achieved precision and recall equivalent to manual annotation.

Technological Inflection Points Several concurrent technological developments make our approach particularly timely:

- LLMs have crossed a capability threshold for reliable structured extraction
- Inference costs are declining rapidly, making large-scale processing economical
- Specialized fine-tuning techniques enable domain-specific performance improvements
- Prediction markets are maturing with more liquidity and expert participation
- Bayesian modeling tools are becoming more accessible and computationally efficient

Together, these developments create a unique opportunity to automate and scale previously manual processes in AI governance modeling.

Neglected Opportunity in Automated Approaches Despite these enabling technologies, automated approaches to AI governance modeling remain surprisingly underexplored:

- Most forecasting platforms still rely entirely on human judgment without model integration
- Governance frameworks rarely incorporate formal causal models
- Technical alignment research and governance discourse remain largely separate
- Few projects are exploring LLM automation for knowledge extraction in this domain

By addressing this neglected opportunity, our project can create substantial leverage in the AI governance ecosystem, dramatically scaling the information processing capacity of the field at a critical moment in AI development.

4. Methodology: Technical Implementation and Validation

4.1 System Architecture and Data Flow

Our system architecture implements an end-to-end pipeline from unstructured text to actionable insights. The architecture consists of five interconnected components, each handling a specific aspect of the workflow:

Text Ingestion and Preprocessing

- Source documents (papers, blog posts, expert reports) enter the system through APIs or manual upload
- Documents undergo preprocessing including format normalization, metadata extraction, and relevance filtering
- Preprocessed documents are stored in a version-controlled repository with citation information preserved

LLM-Powered Extraction Layer

- Documents are analyzed using a two-stage process:
    1. Identification of key variables, claims, and uncertainty expressions
    2. Mapping of relationships between identified elements
- Extraction occurs through carefully engineered prompts to frontier LLMs (Claude, GPT-4)
- Extracted structures are represented in an intermediate ArgDown format (a markdown-like notation for structured argument mapping, documentation at: https://argdown.org) that captures argument structure with syntax like `<premises>` and `[hypothesis]`

Bayesian Network Construction Module

- ArgDown representations are transformed into formal Bayesian networks
- Nodes represent variables identified in the extraction phase
- Edges represent causal relationships and dependencies
- Conditional probability tables are populated based on extracted probability judgments
- When explicit probabilities are absent, LLMs generate estimates based on contextual reasoning

Forecasting Integration Layer

- External forecasting data is ingested through APIs from platforms like Metaculus
- Forecasts are mapped to corresponding variables in the Bayesian network through a combination of semantic matching algorithms and expert-defined mappings
- Weighting algorithms determine the influence of different forecast sources
- Update mechanisms maintain synchronization between forecasts and network parameters

Interactive Visualization and Analysis Interface

- Users interact with the system through a web-based interface
- Visualization components display network structure and probability distributions
- Analysis tools enable query execution, sensitivity analysis, and counterfactual reasoning
- Policy evaluation features support intervention modeling and scenario comparison

Data flows between these components through standardized formats, with metadata tracking the provenance and uncertainty of all information. This architecture ensures that the system remains modular, allowing individual components to be improved independently while maintaining end-to-end functionality.

4.2 Automated Extraction Pipeline

The automated extraction pipeline represents a core technical innovation of our project. It transforms unstructured text into structured knowledge representations through several specialized steps:

ArgDown Intermediate Representation We employ ArgDown—a markdown-like notation for argument mapping developed by Christian Voigt (2014)—as an intermediate representation between natural language and formal Bayesian networks. ArgDown captures:

- Statements (claims about the world, represented as [Statement])
- Premises (supporting evidence or reasoning, represented as <Premise>)
- Support relationships (indicating how premises support statements, represented as =>)
- Attack relationships (indicating rebuttals or counterarguments, represented as =/=>)
- Undercutting relationships (challenging inferential connections rather than conclusions, represented as =|=>)

For example, an argument might be represented as: <AI systems will continue to improve rapidly> => [Advanced AI systems pose existential risk]

This intermediate representation preserves the argumentative structure of source texts while providing sufficient formality for subsequent transformation into Bayesian networks.

Two-Stage Prompting Approach Our extraction uses a two-stage LLM prompting strategy to maximize accuracy:

Stage 1: Identification

- Prompt LLMs to identify all claims, premises, and evidence in the source text
- Extract explicit probability judgments and uncertainty expressions

- Identify conditional statements and counterfactual reasoning

Stage 2: Structuring

- Prompt LLMs to map relationships between the identified elements
- Determine support, attack, and undercutting relationships
- Organize elements into a coherent argument structure

This staged approach outperforms end-to-end extraction in our preliminary experiments when compared to ground truth annotations.

Specialized Handling for Complex Cases The extraction pipeline includes specialized handling for challenging cases frequently encountered in AI safety literature:

- Implicit Premises: LLMs identify unstated assumptions that authors rely on
- Nested Conditionals: Multi-level conditions are preserved through explicit dependency tracking
- Citation-Backed Claims: Evidence from citations is distinguished from direct argumentation
- Quantitative Uncertainty: Numerical probabilities are extracted directly when present
- Qualitative Uncertainty: Linguistic expressions of uncertainty are mapped to probability ranges

For example, when a text states "X is likely to cause Y if Z is present," our system captures both the conditional relationship and the uncertainty expression, mapping "likely" to an appropriate probability range based on calibration studies.

Quality Assurance and Validation The extraction pipeline incorporates several validation mechanisms:

- Multiple extraction runs with different prompts to assess consistency
- Checks for logical and structural coherence in extracted arguments
- Identification of contradictions or circular reasoning
- Comparison of extraction results against a growing database of expert annotations

These validation processes ensure that the extracted structures faithfully represent the source materials while identifying areas where human review may be necessary.

4.3 Bayesian Network Construction and Inference

Once arguments are extracted in ArgDown format, we construct formal Bayesian networks for probabilistic reasoning and inference.

DAG Construction from Extracted Arguments The transformation from ArgDown to DAG follows a systematic procedure:

1. Each statement and premise becomes a node in the graph
2. Support relationships become directed edges from premise to statement
3. Attack relationships become directed edges with negative influence

4. Undercutting relationships modify the strength of existing edges

This transformation preserves the argumentative structure while creating a mathematically tractable representation for probabilistic inference.

Probability Table Population Conditional probability tables (CPTs) are populated through a combination of:

- Direct extraction of explicit probabilities from source texts
- LLM-generated estimates based on contextual cues and strength of arguments
- Expert elicitation for critical parameters
- Prior distributions from related forecasting questions

For nodes with many parents—a common challenge in complex models—we may employ specialized representations like noisy-OR and noisy-AND. These reduce parameter requirements from exponential (2^n for n binary parents) to linear (n parameters) while maintaining representational adequacy for many real-world causal relationships.

Consistency Enforcement and Calibration Mathematical consistency is enforced through:

- Ensuring all probability distributions sum to 1
- Applying Bayes' theorem to validate conditional probabilities
- Checking for coherence across the joint distribution
- Detecting and resolving inconsistencies between different information sources

Additionally, we calibrate the model against known benchmarks where available and perform sensitivity analysis to identify parameters with disproportionate influence on key outputs.

Inference Techniques for Complex Networks For inference in large, complex networks, we employ:

- Variable elimination for exact inference in tractable subnetworks
- Junction tree algorithms for efficient exact inference
- Importance sampling for approximate inference in larger networks
- Markov Chain Monte Carlo methods for complex queries

These techniques balance computational efficiency with accuracy, allowing meaningful analysis even in large models with intricate dependency structures.

4.4 Prediction Market Integration Module

To keep our models current with the latest expert judgments, we integrate with prediction markets and forecasting platforms through specialized connectors.

API Connections with Forecasting Platforms We establish automated connections with:

- Metaculus: Accessing probability distributions for AI-relevant questions (API calls)
- Manifold Markets: Incorporating market-based forecasts for granular questions
- Good Judgment: Integrating superforecaster predictions for high-stakes questions

- Epoch: Tracking real-world AI development against predictions

These connections retrieve both current forecasts and historical data, enabling trend analysis and comparison of forecast evolution over time. Each forecast is mapped to corresponding variables in our Bayesian network through a combination of semantic matching algorithms and expert-defined mappings.

Weighting Mechanisms for Source Reliability Not all forecasts are equally reliable. We implement a weighting system based on:

- Track record of forecasters or platforms
- Relevance of the forecast to the specific variable
- Recency and update frequency
- Consistency with other information sources
- Expert assessment of forecast quality

These weights dynamically adjust based on performance, ensuring that the most reliable sources have greater influence on the model.

Real-Time Update Procedures Our system maintains synchronization between forecasts and model parameters through:

- Scheduled polling of API endpoints at appropriate intervals
- Event-triggered updates when significant forecast changes occur
- Batch processing to incorporate multiple forecast updates efficiently
- Anomaly detection to flag unusual or potentially erroneous forecast movements

This real-time updating ensures that our models reflect the latest information without requiring manual intervention.

4.5 Validation Methodology

Rigorous validation is essential for establishing the credibility of our approach. We employ a multi-faceted validation strategy:

Comparison with Expert Annotations We assess extraction accuracy against a growing database of expert annotations:

- Multiple experts independently analyze source documents
- Inter-annotator agreement establishes a reliability baseline
- Automated extraction is compared against the consensus annotation
- Discrepancies are analyzed to identify systematic errors or biases

This process provides quantitative metrics on precision, recall, and F1 scores for different aspects of extraction.

Ablation Studies to Identify Critical Components We conduct systematic ablation studies to:

- Identify which components contribute most to overall performance

- Measure the impact of different prompting strategies
- Assess the value of the two-stage extraction approach
- Evaluate the contribution of specialized handling for complex cases

These studies guide our development priorities and highlight components requiring additional refinement.

Red-Teaming Approaches We proactively identify failure modes through structured red-teaming:

- Adversarial document creation designed to challenge the extraction system
- Edge case testing with unusual argument structures and reasoning patterns
- Stress testing with exceptionally complex or ambiguous texts
- Cross-domain validation using materials from adjacent fields

This red-teaming approach helps us identify and address vulnerabilities before deployment, improving overall robustness.

5. Expected Outcomes and Applications

5.1 World Model Extraction and Analysis Tool

Our initial concrete deliverable will be a functional World Model Extraction and Analysis Tool—a system that transforms AI safety literature into structured causal models for analysis and evaluation. This tool embodies our broader methodology and demonstrates its practical utility.

Interface Specifications The tool will feature:

- An intuitive web-based interface accessible to technical and non-technical users
- Customizable input panels for adjusting key variables and assumptions
- Visual representation of the underlying Bayesian network
- Real-time updating of probabilities as inputs change
- Ability to save and share specific configurations
- Comparison views for examining different worldviews or scenarios
- Structured feedback mechanisms to capture user insights for iterative improvement

Users will be able to modify assumptions according to their beliefs, enabling exploration of how different premises lead to different conclusions about existential risk.

Visualization Approaches Complex probabilistic models require sophisticated visualization techniques:

- Interactive node-link diagrams showing causal relationships
- Heat maps indicating variable sensitivity and impact
- Tornado diagrams highlighting key uncertainties
- Probability distribution plots for outcomes and critical variables
- Time-series projections showing how risks may evolve

These visualizations make the underlying model accessible and interpretable, converting abstract probabilities into actionable insights.

Sensitivity Analysis and Scenario Exploration The tool will support in-depth exploration through:

- One-at-a-time sensitivity analysis for individual parameters
- Global sensitivity analysis identifying interaction effects
- Scenario definition and comparison utilities
- "Backward reasoning" to identify conditions needed for specific outcomes
- Critical path analysis highlighting necessary and sufficient conditions

These features enable users to identify which uncertainties matter most for their conclusions and where additional research or evidence gathering would be most valuable.

Practical Utility Assessment: With sufficient funding, we will implement a structured framework to evaluate real-world utility, including: (1) targeted case studies with policy stakeholders, (2) comparison against expert consensus, and (3) retrospective analysis of system insights.

5.2 The AI Grand Strategy Framework

Beyond the extraction tool, we will develop a comprehensive framework for AI Grand Strategy that builds on our technical infrastructure.

Strategy Evaluation Across Worldviews The framework will:

- Formalize diverse worldviews from the AI safety community
- Identify robust strategies that perform well across multiple worldviews
- Highlight critical disagreements that drive strategy divergence
- Map conditions under which different strategies become optimal

This approach transcends typical strategy development by explicitly modeling how different assumptions lead to different strategic conclusions.

Criteria for Strategy Evaluation Strategies will be evaluated against multiple criteria:

- Expected risk reduction across probability distributions
- Robustness to uncertainty in key parameters
- Adaptability to changing conditions and new information
- Political and technical feasibility
- Potential for unintended consequences
- Interaction effects with other strategic elements

These criteria ensure that recommended strategies are both theoretically sound and practically implementable.

Adaptation Mechanisms Unlike static strategy documents, our framework will include:

- Explicit conditional branches specifying how strategy should adapt as conditions change

- Triggers for strategy reevaluation based on new information
- Continuous integration of emerging research and forecast updates
- Periodic full reviews to incorporate structural model changes

This adaptive approach ensures that the strategy remains relevant as the AI landscape evolves.

5.3 AGI Risk Monitor Visualization

To communicate AI risk levels to broader audiences, we will develop an AGI Risk Monitor—a visual representation inspired by the Bulletin of Atomic Scientists' Doomsday Clock but specifically focused on AI risks.

Visual Representation of Risk Indicators The monitor will feature:

- An intuitive interface showing proximity to high-risk conditions
- Supplementary indicators for specific risk factors
- Color coding for different risk categories and sources
- Historical tracking showing how risk assessments have changed
- Explanatory components detailing the reasoning behind current assessments

This visualization translates complex risk models into an accessible format for policymakers, journalists, and the public.

Update Mechanisms and Transparency The monitor will maintain credibility through:

- Algorithmic updating based on the underlying Bayesian network
- Clear documentation of how indicators affect the risk assessment
- Transparent methodology for integrating different information sources
- Version history tracking showing when and why assessments changed
- Expert review panels validating major assessment movements

These features ensure that the monitor isn't merely a subjective assessment but a principled reflection of the underlying risk model.

Educational Components Beyond risk communication, the monitor will include:

- Interactive explainers for key AI risk concepts
- User-adjustable controls to explore how different factors affect risk
- Contextual information linking current developments to risk assessments
- Resources for deeper understanding of specific concerns
- Recommendations for relevant research and policy proposals

These educational elements help users develop more sophisticated mental models of AI risk, increasing the overall quality of public discourse.

5.4 Cross-Model Comparison Tools

A unique contribution of our approach is the ability to compare different world models and identify sources of agreement and disagreement.

Techniques for Identifying Agreement and Disagreement Our tools will implement:

- Structural comparison identifying shared and distinct causal pathways
- Parameter comparison highlighting differences in probability estimates
- Outcome analysis showing how models diverge in conclusions
- Critical variable identification focusing on key disagreements

These techniques help stakeholders understand where genuine disagreements exist versus where differences might be merely terminological or superficial.

Visualization of Worldview Differences We will create specialized visualizations for model comparison:

- Side-by-side network displays highlighting structural differences
- Overlay views showing where models agree and disagree
- Difference heat maps indicating parameter divergences
- Outcome distribution comparisons across models

These visualizations make complex model differences immediately apparent, facilitating productive discussions about genuine cruxes.

Consensus Model Construction Where appropriate, our tools will support:

- Automated identification of shared structures across models
- Aggregation of probability estimates from different sources
- Explicit representation of disagreements as probability distributions
- Construction of minimal models that capture essential disagreements

These capabilities help build shared understanding even when complete consensus isn't possible, creating foundations for coordination despite differing perspectives.

5.5 Other Possible Tools

Beyond our core deliverables, we envision several additional tools that could be developed as extensions:

Policy Impact Simulator

- Interactive simulation of policy interventions
- Counterfactual analysis capabilities
- Cost-benefit assessment frameworks
- Multi-objective optimization tools
- Stakeholder impact analysis

AI Progress Tracker

- Benchmarking of AI capabilities against predefined metrics
- Forecast comparison with actual developments
- Early warning indicators for capability jumps
- Integration with technical ML research literature
- Trend analysis and trajectory projection

Coordination Platform

- Shared workspaces for collaborative modeling
- Version control for model development
- Commenting and annotation capabilities
- Expert elicitation protocols
- Consensus-building mechanisms

Strategic Early Warning System

- Real-time monitoring of key indicators
- Anomaly detection for unexpected developments
- Threshold alerts for critical variables
- Escalation protocols for high-risk conditions
- Automated briefing generation for rapid response

While these tools extend beyond our initial scope, they represent natural expansions of our framework and methodology as the project matures.

6. Theory of Change and Impact Assessment

6.1 Direct Impact Pathways

Our project creates impact through several distinct but complementary pathways, each targeting a critical leverage point in the AI governance ecosystem.

Improving Research Allocation By making implicit models explicit and identifying key uncertainties, our tools help researchers prioritize their efforts more effectively. Specifically:

- Technical alignment researchers can focus on aspects most critical for reducing existential risk
- Governance researchers can identify policy levers with the highest expected impact
- Forecasters can target questions with the greatest decision relevance

If this improves research allocation efficiency even modestly across the field, the impact would be substantial given the high stakes and limited research resources.

Enhancing Decision-Making for Policymakers For policymakers navigating complex AI governance questions, our tools provide:

- Structured frameworks for assessing policy impacts across multiple scenarios
- Explicit quantification of uncertainties and their implications

- Comparison of different expert perspectives in a common language
- Identification of robust interventions across worldviews

We will engage directly with policy stakeholders through targeted briefings, workshops, and consultation sessions to ensure these tools inform actual governance decisions.

Facilitating Coordination Across Domains Perhaps most importantly, our tools create a shared epistemic infrastructure that facilitates coordination:

- Technical and governance researchers can more easily communicate through formal models
- Different organizations can align their strategies based on shared understanding
- Researchers across geographic and institutional boundaries can contribute to a common framework
- Distinct philosophical perspectives can be represented within a unified system

This coordination benefit scales superlinearly with adoption, as each additional participant increases the value of the system for all existing users.

6.2 Expected Impact Quantification

While precise quantification of impact remains challenging, we can provide reasonable estimates across several dimensions:

Efficiency Gains in AI Governance Research

- Current estimate: ~500 researchers working on AI governance globally
- Average researcher cost: ~$150,000/year (salary + overhead)
- Potential efficiency improvement through better coordination
- Value: More effectively allocated research effort

Value of Information from Improved Modeling

- Baseline estimates of existential risk from expert surveys
- Potential reduction through better governance
- Value of this reduction: Enormous given the existential stake (8 billion lives + future generations)
- Even with substantial uncertainty, the expected value is orders of magnitude higher than project costs

Strategic Coordination Improvements

- Current coordination challenges (duplicate efforts, misaligned initiatives)
- Expected improvements through shared models
- Value: More coordinated research and advocacy

These estimates focus on immediate, measurable effects rather than long-term impacts on existential risk, which would yield much larger expected value calculations.

6.3 Differential Advancement Considerations

Any project in this domain must carefully consider potential risks of advancing certain types of knowledge or capabilities. We have conducted a thorough differential advancement assessment:

Information Hazards and Mitigation Strategies Potential hazards:

- Revealing vulnerabilities in existing governance approaches
- Providing optimization targets for actors seeking to evade governance
- Creating false confidence in flawed models
- Accelerating technical capabilities through certain types of analysis

Mitigation strategies:

- Pre-publication review by information security experts
- Graduated access to sensitive analytic capabilities
- Focus on defensive applications rather than exploits
- Explicit uncertainty representation to prevent overconfidence
- Emphasis on governance rather than technical capabilities

Data Governance and Privacy We will establish clear protocols for handling potentially sensitive insights extracted from world models, including anonymization where appropriate, tiered access controls for different stakeholder groups, and explicit policies regarding what information will be made public versus restricted. We will try to avoid handling sensitive data altogether, especially in the beginning phases of the project.

Safeguards Against Misuse We will implement:

- An ethics review process for all publications and tools
- Terms of use prohibiting harmful applications
- Monitoring for potential misuse of public tools
- Technical safeguards preventing certain types of analysis
- Consultation with security experts on release decisions

Sensitive Forecast Handling For particularly sensitive forecasts:

- Implement access controls based on need-to-know
- Apply differential privacy techniques where appropriate
- Aggregate information to prevent reverse engineering
- Create secure environments for sensitive analysis
- Establish clear guidelines for what information should be public vs. restricted

Our general approach favors openness where possible, with restrictions only when specific, articulated risks outweigh the benefits of transparency.

6.4 Dissemination Strategy

Effective dissemination is critical for translating our research into real-world impact. Our strategy targets multiple channels:

Publishing Approach We will disseminate our work through:

- Academic papers in AI safety, ML, and governance venues
- EA and rationality forums (LessWrong, Alignment Forum, EA Forum)
- Interactive web platforms hosting our tools
- Policy briefs tailored to specific stakeholder groups
- Technical documentation for researchers and developers

Each publication will be adapted to its specific audience while maintaining consistency in the underlying models and analyses.

Stakeholder-Specific Communication We will develop targeted communications for:

- Technical AI safety researchers: Emphasizing formal model details and validation
- Governance researchers: Focusing on policy implications and intervention assessment
- Policymakers: Highlighting actionable insights and decision support
- General public: Providing accessible explanations of key concepts and concerns
- Adjacent communities: Connecting our work to related fields and approaches

This tailored approach ensures that our research reaches and influences the most relevant audiences.

Community Building Initiatives Beyond publications, we will:

- Host workshops on using our tools for research and analysis
- Create tutorial materials for incorporating our approaches into existing workflows
- Establish working groups focused on specific applications
- Organize collaborative modeling sessions across organizations
- Develop an online community around model development and refinement

These initiatives build the human infrastructure necessary for long-term impact, creating a community of practice around our methodological approach.

7. Implementation Timeline and Risk Management

7.1 Phase-Based Development Approach

Our implementation follows a carefully structured timeline with distinct phases, each building on previous work while allowing for parallel development of connected components.

Phase 1: Foundation Development (Months 1-4)

- Comprehensive literature review and stakeholder interviews (Month 1)
- Technical infrastructure setup and Bayesian network design (Month 1-2)
- Initial extraction system prototype development (Month 2-3)
- Worldview extraction experiments (Month 3-4)

- Internal testing and refinement (Month 4)

Major Milestone: Working World Model Extraction Tool Prototype (End of Month 4)

Phase 2: Core Tool Development (Months 5-8)

- Expert feedback collection and incorporation (Month 5)
- Worldview extraction system development (Month 5-6)
- Prediction market API integration (Month 6-7)
- Policy impact evaluation module development (Month 7-8)
- Integration testing of all components (Month 8)

Major Milestone: Integrated Tool Suite (End of Month 8)

Phase 3: Scaling and Strategy Development (Months 9-12)

- Public beta release and community testing (Month 9-10)
- Automated world model extraction at scale (Month 10-11)
- Strategic pattern identification (Month 10-11)
- AI Grand Strategy framework development (Month 11-12)
- Documentation and knowledge transfer (Month 12)

Major Milestone: AI Grand Strategy Framework (End of Month 12)

This timeline includes ~20% buffer time distributed across phases (with more buffer allocated to phases with higher uncertainty, particularly the automated extraction system development) to account for unexpected challenges and ensure quality deliverables. We've structured the development so that each phase produces valuable outputs even if subsequent phases encounter difficulties. The core team will be fully engaged throughout all phases, with Valentin Meyer leading the technical implementation aspects and Coleman Snell focusing on stakeholder engagement, validation, and strategic deployment.

7.2 Key Risks and Mitigation Strategies

We have identified several key risks to project success and developed specific mitigation strategies for each:

Technical Risk: LLM Extraction Quality Insufficient

- Risk level: Medium-High
- Impact: Could significantly reduce the scale and accuracy of world model extraction
- Mitigation:
    - Develop hybrid human-AI approaches that leverage LLMs while incorporating human oversight
    - Create structured templates to guide extraction and reduce ambiguity
    - Build a modular system where human input can substitute for automation where necessary

    - Continuously measure extraction quality and focus development on identified weaknesses

Coordination Risk: Stakeholder Engagement Limitations

- Risk level: Medium
- Impact: Could reduce the uptake and influence of developed tools
- Mitigation:
    - Conduct early user research to understand stakeholder needs and expectations
    - Create multiple interfaces with varying complexity for different user groups
    - Demonstrate concrete value through case studies of tool applications
    - Establish an advisory group of potential users to guide development priorities

Resource Risk: Computational Requirements Exceed Budget

- Risk level: Medium
- Impact: Could limit the scale of model development and analysis
- Mitigation:
    - Implement efficient algorithms that minimize computational requirements
    - Develop hierarchical approaches that allow focused analysis of critical subnetworks
    - Establish cloud computing partnerships or academic computing resource access
    - Prepare scaled-back implementations that preserve core functionality with reduced resources

Epistemic Risk: Model Uncertainty Handling Challenges

- Risk level: Medium-High
- Impact: Could produce misleading results if uncertainty is inadequately represented
- Mitigation:
    - Implement explicit uncertainty quantification for all model components
    - Develop sensitivity analysis tools to identify key uncertainties
    - Incorporate multiple expert perspectives to capture disagreement
    - Maintain transparency about limitations and simplifying assumptions

Scope Risk: Feature Creep and Expansion

- Risk level: High
- Impact: Could dilute focus and prevent completion of core deliverables
- Mitigation:
    - Implement strict scope management with explicit decision criteria
    - Prioritize core deliverables with clear definitions of completion
    - Create modular architecture where extensions can be developed independently
    - Establish a change control process requiring justification for scope modifications

7.3 Early Stopping Criteria and Pivot Points

We recognize that research often reveals unexpected challenges or opportunities. We've established clear criteria for when to consider early stopping or pivoting:

After World Model Extraction Tool Development (Month 4)

- Stop criteria: Feedback indicates limited additional value from scaling, or technical challenges prove more significant than anticipated
- Deliverable: Working extraction tool with documentation and examples
- Impact: Still provides valuable probabilistic modeling infrastructure

After Initial Worldview Extraction System (Month 6)

- Stop criteria: Automation proves fundamentally limited, or more promising approaches emerge elsewhere
- Deliverable: Validated extraction methodology with examples and limitations documentation
- Impact: Advances knowledge about LLM capabilities for knowledge extraction

After Policy Evaluation Framework (Month 8)

- Stop criteria: Full strategy development faces insurmountable challenges or higher-impact opportunities emerge
- Deliverable: Framework for evaluating policy interventions with case studies
- Impact: Provides valuable decision support tools for policymakers

Each stopping point would still yield publishable results and useful tools, ensuring that research effort translates to value even if the full project scope proves unachievable.

8. Team Composition and Capabilities

8.1 Core Team Expertise

Our team brings together precisely the interdisciplinary expertise needed for this complex project:

Valentin Jakob Meyer

- **Expertise**: Bayesian networks, probabilistic modeling, epistemology, forecasting theory
- **Experience**: Extensive work implementing and analyzing Bayesian networks for complex decision problems, including directed acyclic graphs (DAGs) and probabilistic graphical models
- **Skills**: Mathematical modeling, causal inference, uncertainty quantification
- **Relevance**: These capabilities are directly applicable to the formal modeling framework at the heart of our approach

Coleman Snell

- **Expertise**: AI governance, ethics, strategic planning, community building
- **Experience**: Research at AI:FAR, University of Chicago's X Risk Lab, and Cambridge's Center for the Study of Existential Risk (CSER)
- **Skills**: Stakeholder engagement, policy analysis, science communication

- **Relevance**: Ensures our technical tools connect to real-world governance needs and stakeholder requirements

Together, our complementary backgrounds create a uniquely qualified team:

- We bridge technical modeling and governance domains—a rare combination essential for this project
- We have demonstrated ability to communicate complex technical concepts to diverse audiences
- Our combined networks span the AI safety, governance, and forecasting communities
- We each bring deep domain knowledge in complementary aspects of the project

Previous Relevant Projects:

- Conducted manual worldview extraction from influential AI safety papers
- Developed prototype Bayesian network models for specific AI risk pathways
- Published analyses of AI governance approaches in respected forums
- Successfully implemented API integrations with prediction platforms
- Designed and executed expert elicitation protocols for uncertainty quantification

These experiences provide a strong foundation for the proposed work, demonstrating our ability to execute on both technical and strategic aspects of the project.

8.2 Advisory Network and Collaborations

Our work doesn't exist in isolation. We've established an advisory network and collaborations that strengthen our approach:

Key Advisors

- **Names**: Matthew Genzel, Sean Ó hÉigeartaigh, Johanne Meyer, Thomas Porter, and more names to be added later

- **Technical advisors**: Experts in probabilistic modeling, causal inference, and LLMs

- **Domain advisors**: Specialists in AI safety, governance, and forecasting

- **Implementation advisors**: Professionals with experience developing similar tools

Organizational Collaborations

- MTAIR team members providing technical guidance and knowledge transfer
- Forecasting platforms for data access and integration support
- AI safety research organizations for domain expertise and validation
- Academic institutions for computational resources and peer review

Collaboration Mechanisms

- Regular advisory board meetings for project oversight
- Technical working groups for specific challenges
- User testing panels for interface development
- Expert elicitation protocols for model parameterization

These collaborations multiply our impact by leveraging the expertise and resources of the broader ecosystem while ensuring our work remains connected to complementary efforts.

8.3 Funding Efficiency

We've designed this project to maximize impact per dollar through several efficiency mechanisms:

Resource Allocation Optimization

- Focus technical resources on automation components with highest leverage
- Utilize existing open-source tools and libraries where possible
- Implement cloud-based architecture to minimize infrastructure costs
- Deploy graduated development allowing early value creation

Comparative Impact Analysis Our expected impact per dollar compares favorably with alternatives:

- Manual modeling approaches require ~5-10x more person-hours for similar coverage
- Dedicated forecasting projects lack integration with causal models
- Policy analysis without formal modeling provides less decision support
- Technical AI safety work without governance connection lacks implementation pathways

Leverage Points for Outsized Returns We've identified specific leverage points where funding creates disproportionate impact:

- LLM automation creates scaling effects that multiply human capacity
- Tool development generates ongoing value beyond the project timeframe
- Integration across domains creates network effects in coordination
- Strategic framework development enables alignment of much larger resources

By focusing on these high-leverage opportunities, we maximize the return on the requested funding while creating sustainable value.

9. Resource Requirements and Allocation

9.1 Budget Breakdown and Justification

Our project requires the following resources over its 12-month timeline:

Personnel Costs (70% of total budget)

- Core team compensation (2 FTE)
    - Includes salary, benefits, and taxes for two full-time researchers

    - Allocation: 40% technical development, 40% research, 20% stakeholder engagement
- Expert consultations
    - Funds specialized expertise for validation, review, and technical advice
    - Allocation: 15-20 expert-days across the project timeline

Technical Infrastructure (15% of total budget)

- LLM API access
    - Covers costs for GPT-4, Claude, and other frontier model access
    - Estimated usage: ~10M tokens per month for extraction and analysis
- Cloud computing resources
    - Supports computation for complex Bayesian networks and simulations
    - Includes development, staging, and production environments
- Software and tools
    - Licenses for specialized modeling software (e.g., Analytica)
    - Development tools and services

Community Engagement and Dissemination (10% of total budget)

- Workshop and event organization
    - Facilitates expert engagement, feedback collection, and result dissemination
    - Includes virtual and in-person components
- User interface design
    - Professional design services for key user interfaces
    - Ensures tools are accessible to target audiences
- Publication and documentation
    - Covers production of research papers, technical documentation, and policy briefs
    - Includes editing, visualization, and distribution costs

Contingency and Flexibility (5% of total budget)

- Reserved for addressing unexpected challenges or opportunities
- Enables rapid response to emerging needs or promising directions

This budget has been carefully optimized to balance necessary resources with efficient allocation, ensuring every dollar directly contributes to project objectives.

9.2 Alternative Funding Scenarios

We recognize that funding availability may vary. We've developed alternative scenarios to accommodate different resource levels:

Minimal Viable Funding

- Core deliverables:
    - Basic extraction tool with manual assistance
    - Limited automation proof-of-concept
    - Simplified policy evaluation framework

    - 1-2 case studies demonstrating utility
- Implementation approach:
    - Focus on methodological development over automation
    - Reduce scale of model complexity and coverage
    - Leverage more volunteer and community contributions
    - Prioritize academic outputs over tool development

Optimal Funding

- Core deliverables:
    - Comprehensive extraction system with full automation
    - Integrated forecasting platform connections
    - Complete policy evaluation framework
    - AI Grand Strategy with robust implementation pathways
    - Public-facing visualization tools
- Implementation approach:
    - Fully develop all automation components
    - Maximize model coverage and complexity
    - Implement comprehensive validation protocols
    - Develop polished, user-friendly interfaces
    - Execute full stakeholder engagement strategy

Expansion Opportunities With additional resources, we could extend the project to include:

- Fine-tuning of LLMs specifically for causal extraction
- Development of an open API for broader ecosystem integration
- Creation of specialized tools for specific stakeholder groups
- Establishment of a permanent update and maintenance infrastructure
- International expansion with multi-language support

These scenarios demonstrate our ability to adapt the project scope based on available resources while maintaining core impact pathways.

9.3 Post-Grant Sustainability

We have developed a strategy for ensuring the project's impact continues beyond the initial grant period:

Long-term Maintenance and Updates

- Open-source core components to enable community maintenance
- Establish update protocols that can be executed with minimal resources
- Create documentation enabling others to extend and adapt the tools
- Design modular architecture allowing independent component evolution

Institutional Adoption Pathways

- Identify host organizations for long-term tool hosting

- Develop transition plans for operational responsibility
- Create institutional partnerships for ongoing development
- Secure commitments for continued engagement from key stakeholders

Open-Source Community Development

- Build a contributor community during the project
- Establish governance structures for community maintenance
- Create contribution guidelines and documentation
- Develop mentorship programs for new contributors

Potential Follow-on Funding Sources

- Identify specific outputs with dedicated funding opportunities
- Prepare applications for sustainability grants
- Explore institutional support from beneficiary organizations
- Consider minimal service models for specialized applications

Through these approaches, we ensure that the value created during the grant period continues to benefit the AI governance ecosystem long after the project formally concludes.

We believe this proposal presents a compelling case for supporting the AMTAIR project. By addressing the critical coordination failure in AI governance through automated knowledge extraction and formalization, we can significantly enhance humanity's ability to navigate the unprecedented challenges posed by advanced AI systems. The tools and frameworks we develop will not only improve strategic decision-making but also create the epistemic infrastructure necessary for effective coordination at a critical moment in AI development.

We welcome any opportunities to discuss this proposal in more detail and address any questions or concerns you may have.