Detecting conflicts between AI risks and company missions: A global analysis using LLM and RAG architecture

Author: Cristian Camilo Herreño Mojica | **UCL Geography – Msc Social and Geographic Data Science**

UCL supervisors: Anwar Musah, Stephen Law | **Sponsor supervisors:** Ke Zhou, Vibhor Agarwal, Daniele Quercia.

1.Introduction

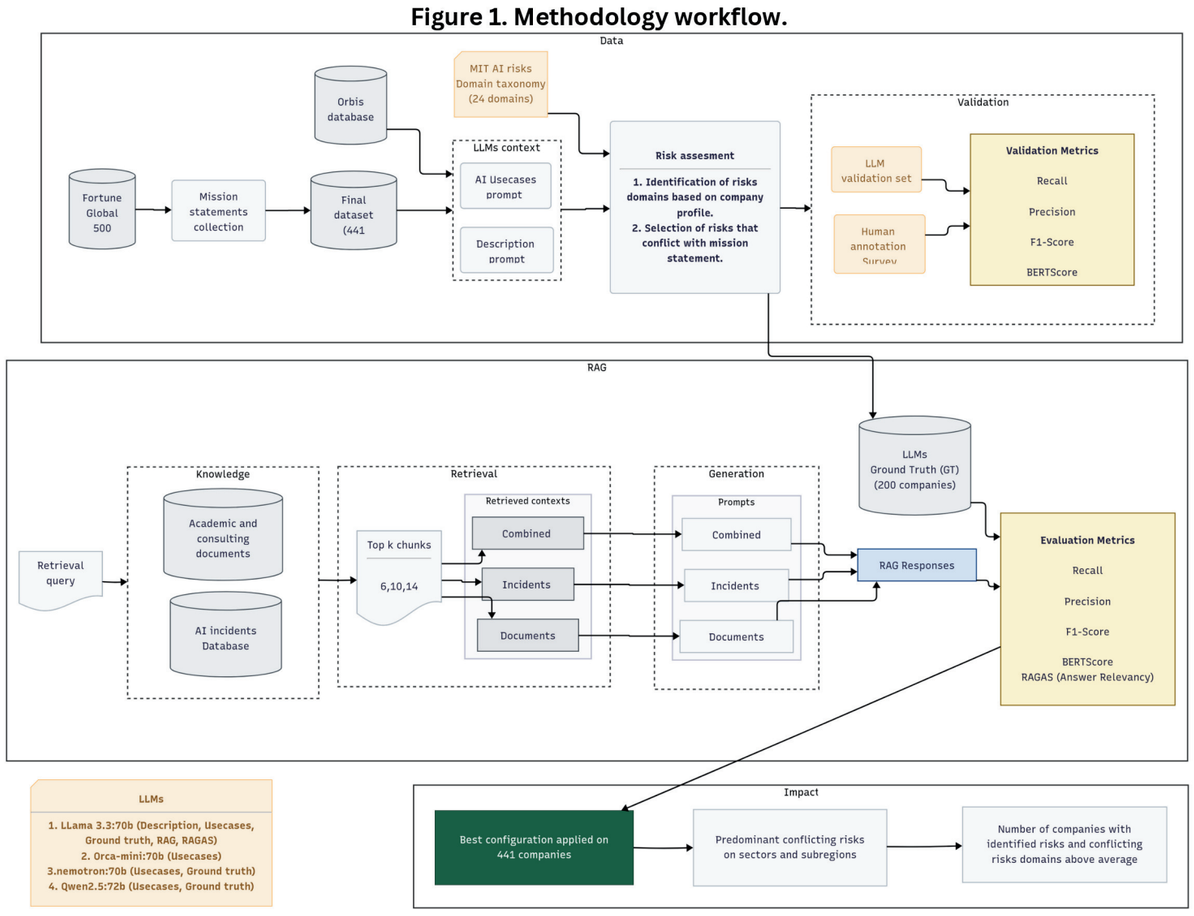
Artificial intelligence (AI) is now embedded in many organisations, driving progress while introducing complex risks. Following Slattery et al. (2024), AI risk is defined as "the possibility of an unfortunate occurrence associated with the development or deployment of AI". Such risks may stem from training data, algorithms, or human interaction, and can cause harm to individuals, organisations, or society.

The NIST AI Risk Management Framework calls for governance that aligns AI systems with organisational values (NIST, 2023). However, practical methods for detecting conflicts between AI risks and these values remain limited. Mission statements, which articulate an organisation's purpose, commitments, and guiding principles, provide a clear reference point for evaluating such conflicts.

Recent advances in Retrieval Augmented Generation (RAG) enable large-scale AI risk analysis (Rao et al., 2025). This dissertation applies RAG to the Fortune Global 500 to systematically identify AI risks, assess their conflict with stated missions, and reveal sector and geographic patterns.

2. Data and methods

Figure 1. Methodology workflow.



Mission statements from 441 companies in the 2024 Fortune Global 500, were collected through automated extraction and manual review. An ensemble of large language models (LLM) generated company descriptions and AI applications, then created a ground-truth dataset for 200 companies by identifying relevant AI risk domains and assessing conflicts with missions. A human-annotated survey of 24 companies, following the same process, provided a set for validation. The main method was RAG, which grounds model outputs in retrieved evidence from a curated knowledge base. Retrieval setups varied by source (sector-specific documents, AI incidents, or both) and chunk size (6, 10, 14).

Data sources:

- **2024 Fortune Global 500:** 500 companies, 441 mission statements collected.
- **2025 Orbis database:** Company profiles and background
- **MIT domain taxonomy:** It classifies AI risks into 7 domains providing a framework for identification.

Knowledge base

- **Documents**: 63 sector-specific documents (3 per sector) from academic and consulting sources
- **AI incidents:** 867 incidents classified under the MIT taxonomy, developed by McGregor (2021)

3.Results

- Across the validation set of 24 companies, the LLM ensemble achieved an F1-score of 0.5 on identifying conflicting risks.
- The optimal configuration tested against the LLM ground truth was the RAG pipeline using a documents-only with k=10 retrieved chunks, achieving the best performance on all metrics.
- This optimal configuration was then applied to analyse the full dataset of 441 companies.
- On average, companies had between 2 and 3 AI risk domains that conflict with their mission statements. Therefore, 4 categories based on the number of risk domains were created (0, 1-3, 4-5,6-7)

Figure 2. RAG results against LLM ensemble for each configuration.

Top k	Evaluation	Metric	Retrieval Source: Documents	Incidents	Combination
k=6	Domains selection	Precision	0.94	0.94	**0.96**
		Recall	**0.58**	0.29	0.30
		F1-score	**0.72**	0.44	0.46
	Mission conflict	Precision	0.91	0.94	**0.96**
		Recall	**0.44**	0.18	0.16
		F1-score	**0.60**	0.30	0.27
	Open text risks	BERTScore (F1)	**0.57**	0.50	0.51
	Conflict explanation	BERTScore (F1)	0.57	**0.53**	**0.53**
	Valid responses	Response rate (%)	**100.00**	**100.00**	**100.00**
	RAGAS	Answer relevancy	**0.43**	0.38	0.39
k=10	Domains selection	Precision	**0.94**	0.92	**0.94**
		Recall	**0.62**	0.35	0.29
		F1-score	**0.74**	0.51	0.45
	Mission conflict	Precision	0.90	0.92	**0.95**
		Recall	**0.48**	0.25	0.20
		F1-score	**0.63**	0.39	0.33
	Open text risks	BERTScore (F1)	**0.58**	0.52	0.52
	Conflict explanation	BERTScore (F1)	**0.59**	0.54	0.53
	Valid responses	Response rate (%)	94.00	**100.00**	**100.00**
	RAGAS	Answer relevancy	**0.46**	0.38	0.40
k=14	Domains selection	Precision	**0.92**	0.87	**0.92**
		Recall	0.15	**0.38**	0.35
		F1-score	0.25	**0.53**	0.50
	Mission conflict	Precision	0.92	0.87	**0.96**
		Recall	0.10	**0.26**	0.21
		F1-score	0.18	**0.40**	0.35
	Open text risks	BERTScore (F1)	**0.58**	0.53	0.52
	Conflict explanation	BERTScore (F1)	0.58	0.54	**0.55**
	Valid responses	Response rate (%)	18.00	**100.00**	98.50
	RAGAS	Answer relevancy	**0.45**	0.39	0.39

References

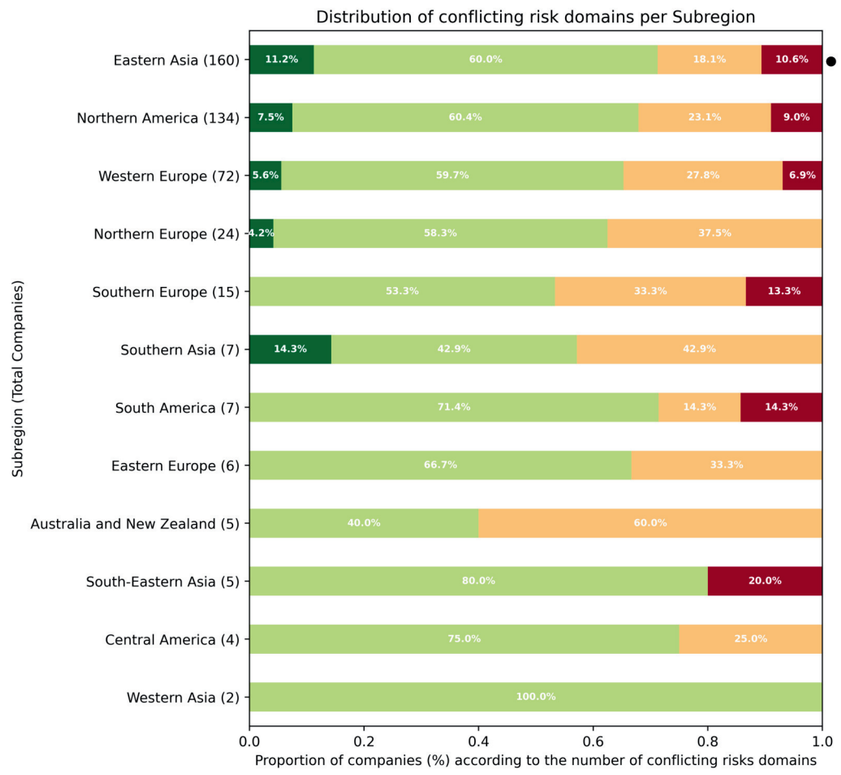
- MCGREGOR, S. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 15458-15463.
- NIST 2023. Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai, 100-1.
- RAO, P. S. B., ŠĆEPANOVIĆ, S., ZHOU, K., BOGUCKA, E. P. & QUERCIA, D. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery.
- SLATTERY, P., SAERI, A. K., GRUNDY, E. A. C., GRAHAM, J., NOETEL, M., UUK, R., DAO, J., POUR, S., CASPER, S. & THOMPSON, N. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. ArXiv, abs/2408.12622.

4. Key findings

Figure 3. Qualitative example of RAG output (Only documents k=10).

Company	Sector	Mission statement	Risk domain & Risks excerpts	Mission conflict explanation
Chubb	Financials	*To be a leader in the global insurance market by always conducting business in a consistent, disciplined and focused manner. Provides "Superior Client Value, Superior Shareholder Value, Superior Employee Value".*	*Privacy & Security:* AI models leak sensitive customer information; Cyberattackers manipulate AI models to extract confidential data.	'Superior Client Value' is conflicted as leaked customer info harms clients, resulting in loss of trust and business.
Repsol	Energy	*An energy company committed to a sustainable world.*	*Socioeconomic & Environmental:* Unintended environmental harm from AI-optimized energy production...	An energy company committed to a sustainable world, yet AI may prioritize efficiency over sustainability, causing environmental harm.

Figure 4. Proportion of companies by the number of conflicting risks domains .



- **Regional**: Australia and S. Europe patterns are shaped by the dominance of Financial and Energy firms. In N. America, customer-centric missions are most challenged by safety and privacy risks, underscoring trust and reliability concerns. In E. Asia, societal-centric missions are most conflicted by socioeconomic and environmental, pointing to potential disruption of social commitments.

- **Sectoral**: Creative sectors face risks across most AI domains, exposing the vulnerability of their aspirational missions. Financial & Healthcare often conflict on safety and privacy due to personal data use. Machinery-intensive sectors mainly conflict with socioeconomic and environmental risks highlighting pressure on sustainability commitments.

5.Conclusion

This study demonstrates that LLM and RAG enables scalable detection of conflicts between organisational missions and AI risks. The results highlight sector and subregion patterns that can inform targeted governance and guide the integration of mission alignment into AI risk assesments.