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 Existential Safety

Valentin Jakob Meyer Prof. Dr. Timo Speith

2025-05-26

The coordination crisis in AI governance presents a paradoxical challenge: unprecedented investment in AI safety coexists alongside fundamental coordination failures across technical, policy, and ethical domains. These divisions systematically increase existential risk by creating safety gaps, misallocating resources, and fostering inconsistent approaches to interdependent problems. This thesis introduces AMTAIR (Automating Transformative AI Risk Modeling), a computational approach that addresses this coordination failure by automating the extraction of probabilistic world models from AI safety literature using frontier language models.

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

```
# [ ]: Configure the Quarto Manuscript options: https://quarto.org/docs/manuscripts/componen
```

1

2 Frontmatter

3 Prefatory Apparatus: Illustrations and Terminology — Quick References

3.1 List of Tables

3.2 List of Graphics & Figures

3.3 List of Abbreviations

esp. especially

f., ff. following

incl. including

p., pp. page(s)

MAD Mutually Assured Destruction

3.4 Glossary

4

5 Introduction

10% of Grade:

• introduces and motivates the core question or problem • provides context for discussion (places issue within a larger debate or sphere of relevance) • states precise thesis or position the author will argue for • provides roadmap indicating structure and key content points of the essay

~ 14% of text ~ 4200 words

- introduces and motivates the core question or problem

5.1 Motivation: Problem Statement

5.2 Motivation: Research Question

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

5.3 Scope: Aim & Context of the Research

5.4 Significance of the Research: Theory of Change

- states precise thesis or position the author will argue for

5.5 Thesis Statement & Position: (Aim of the Paper)

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

5.6 Overview: Structure & Approach of the Paper (Roadmap — Theory of Change)

5.7 Table of Contents

6 Context

20% of Grade:

• demonstrates understanding of all relevant core concepts • explains why the question/thesis/problem is relevant in student's own words (supported by quotations) • situates it within the debate/course material • reconstructs selected arguments and identifies relevant assumptions • describes additional relevant material that has been consulted and integrates it with the course material as well as the research question/thesis/problem

~ 29% of text ~ 8700 words

1. successively (chunk my chunk) introduce concepts/ideas — and 2. ground each with existing literature

7 AMTAIR

20% of Grade:

• provides critical or constructive evaluation of positions introduced • develops strong (plausible) argument in support of author's own position/thesis • argument draws on relevant course material • claim/argument demonstrates understanding of the course materials incl. key arguments and core concepts within the debate • claim/argument is original or insightful, possibly even presents an original contribution to the debate

~ 29% of text ~ 8700 words

8 Discussion

10% of Grade:

• discusses a specific objection to student's own argument • provides a convincing reply that bolsters or refines the main argument • relates to or extends beyond materials/arguments covered in class

~ 14% of text ~ 4200 words

9 Conclusion

10% of Grade:

• summarizes thesis and line of argument • outlines possible implications • notes outstanding issues / limitations of discussion • points to avenues for further research • overall conclusion is in line with introduction

~ 14% of text ~ 4200 words

Bibliography/References

Knuth, Donald E. 1984. "Literate Programming." *Computer Journal* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Marrero, José, Alicia García, Manuel Berrocoso, Ángeles Llinares, Antonio Rodríguez-Losada, and R. Ortiz. 2019. "Strategies for the Development of Volcanic Hazard Maps in Monogenetic Volcanic Fields: The Example of La Palma (Canary Islands)." *Journal of Applied Volcanology* 8 (July). https://doi.org/10.1186/s13617-019-0085-5.

10 Appendices

11 Appendix A

12 Appendix B

13 Appendix C

14 Appendix D

TestText

15 Affidavit

16 Notebooks

17 Quarto Example

17.1 Introduction

```r
eruptions <- c(1492, 1585, 1646, 1677, 1712, 1949, 1971, 2021)
n_eruptions <- length(eruptions)
```

```r
par(mar = c(3, 1, 1, 1) + 0.1)
plot(eruptions, rep(0, n_eruptions),
  pch = "|", axes = FALSE)
axis(1)
box()
```

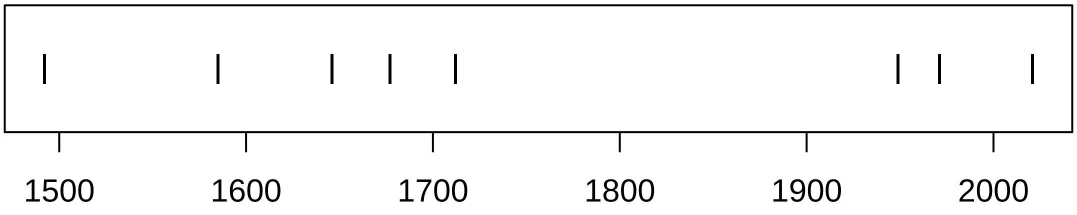

Figure 1: Timeline of recent earthquakes on La Palma

```r
avg_years_between_eruptions <- mean(diff(eruptions[-n_eruptions]))
avg_years_between_eruptions
```

Based on data up to and including 1971, eruptions on La Palma happen every 79.8 years on average.

Studies of the magma systems feeding the volcano, such as Marrero et al. (2019), have proposed that there are two main magma reservoirs feeding the Cumbre Vieja volcano; one in the mantle (30-40km depth) which charges and in turn feeds a shallower crustal reservoir (10-20km depth).

Eight eruptions have been recorded since the late 1400s (Figure 1).

Data and methods are discussed in Section 17.2.

Let x denote the number of eruptions in a year. Then, x can be modeled by a Poisson distribution

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{1}$$

where λ is the rate of eruptions per year. Using Equation 1, the probability of an eruption in the next t years can be calculated.

Table 1: Recent historic eruptions on La Palma

Name	Year
Current	2021
Teneguía	1971
Nambroque	1949
El Charco	1712
Volcán San Antonio	1677
Volcán San Martin	1646
Tajuya near El Paso	1585
Montaña Quemada	1492

Table 1 summarises the eruptions recorded since the colonization of the islands by Europeans in the late 1400s.

La Palma is one of the west most islands in the Volcanic Archipelago of the Canary Islands (Figure 2).

Figure 3 shows the location of recent Earthquakes on La Palma.

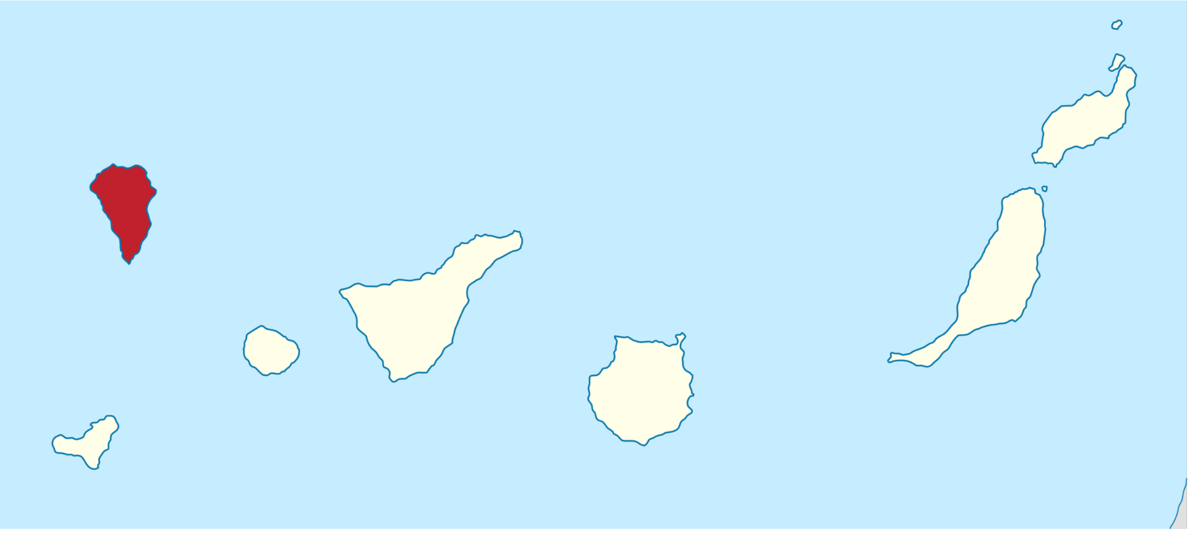

Figure 2: Map of La Palma

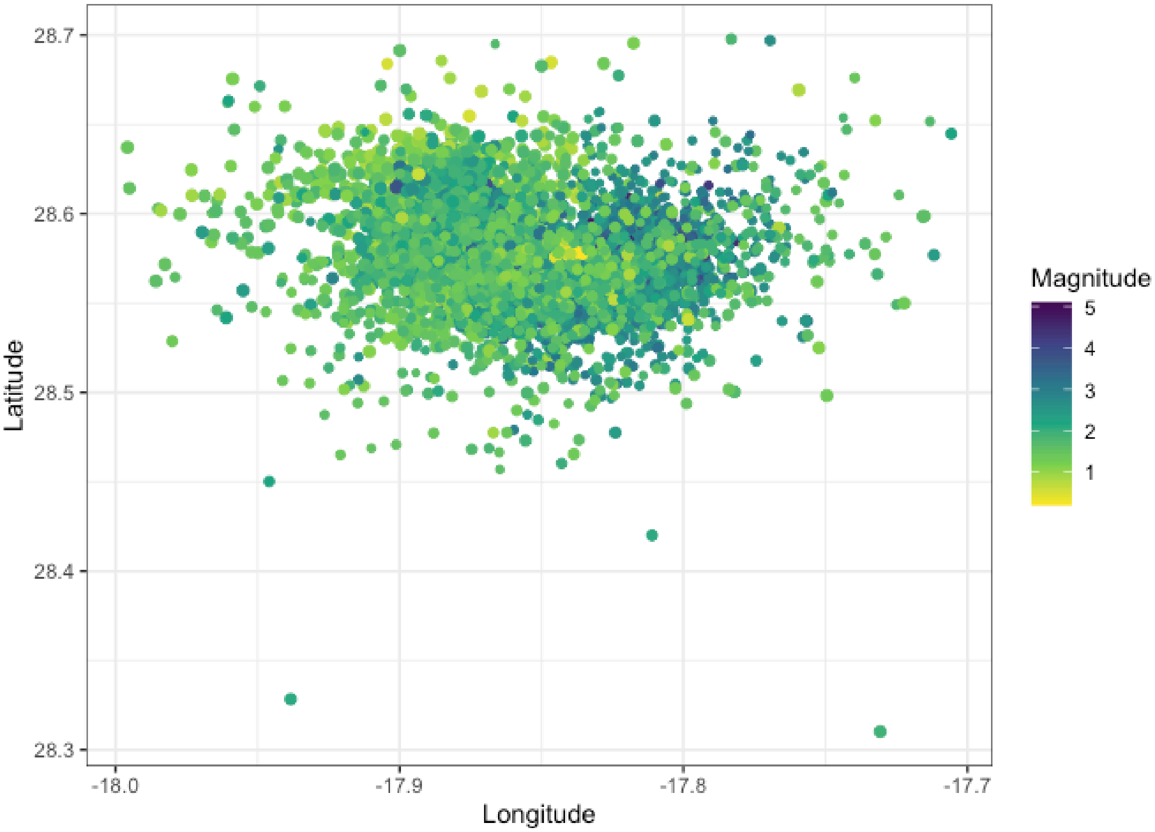

Figure 3: Locations of earthquakes on La Palma since 2017

17.2 Data & Methods

17.3 Conclusion

18 Introduction

This is a booooook created from markdown and executable code.

See [Knuth (1984)] and Knuth (1984) for additional discussion of literate programming.

Regular markdown and $E = mc^2$ equations.

18.1 Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

18.2 Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
2
```

You can add options to executable code like this

```
4
```

The `echo: false` option disables the printing of code (only output is displayed).

```
2
```

More markdown.

18.3 ToDo's

// Double slash creates a new task