

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: Supervisor:

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

Pielmühler Straße 15 52066 Lappersdorf Word Count:
30.000
Source / Identifier:
Document URL

Dr. Timo Speith

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

4 Preface

Abstract

6 Abstract

Outline(s): Table of Contents

Frontmatter

Prefatory Apparatus: Illustrations and Terminology — Quick References

List of Tables

Table 1: Table name
Table 2: Table name
Table 3: Table name

List of Graphics & Figures

List of Abbreviations

esp. especially
f., ff. following
incl. including
p., pp. page(s)
MAD Mutually Assured Destruction

Glossary

Introduction

1.1 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
 - $\sim 14\%$ of text ~ 4200 words
 - introduces and motivates the core question or problem

1.2 Motivation: Problem Statement

1.3 Motivation: Research Question

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

1.4 Scope: Aim & Context of the Research

1.5 Significance of the Research: Theory of Change

• states precise thesis or position the author will argue for

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

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

- 1.7 Overview: Structure & Approach of the Paper (Roadmap Theory of Change)
- 1.8 Table of Contents

Context

2.1 Context

20% of Grade:

- \bullet demonstrates understanding of all relevant core concepts \bullet explains why the question/thesis/problem is relevant in student's own words (supported by quotations) \bullet situates it within the debate/course material \bullet reconstructs selected arguments and identifies relevant assumptions \bullet describes additional relevant material that has been consulted and integrates it with the course material as well as the research question/thesis/problem
 - $\sim 29\%$ of text ~ 8700 words
 - 1. successively (chunk my chunk) introduce concepts/ideas and 2. ground each with existing literature

AMTAIR

3.1 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

 $\sim 29\%$ of text ~ 8700 words

Discussion

4.1 Discussion

10% of Grade:

- \bullet discusses a specific objection to student's own argument \bullet provides a convincing reply that bolsters or refines the main argument \bullet relates to or extends beyond materials/arguments covered in class
 - $\sim 14\%$ of text ~ 4200 words

Conclusion

5.1 Conclusion

10% of Grade:

- \bullet summarizes thesis and line of argument \bullet outlines possible implications \bullet notes outstanding issues / limitations of discussion \bullet points to avenues for further research \bullet overall conclusion is in line with introduction
 - $\sim 14\%$ of text ~ 4200 words

References

24 References

Appendix A

Appendices

- A.1 Appendices
- A.2 Appendix A
- A.3 Appendix B
- A.4 Appendix C
- A.5 Appendix D

 ${\rm TestText}$

A.6 Affidavit

$\begin{array}{c} \mathbf{Appendix} \ \mathbf{B} \\ \mathbf{appendix} \mathbf{A} \end{array}$

testtext