PUCRS

ESCOLA DE HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

NICHOLAS KLUGE CORRÊA

DYNAMIC NORMATIVITY
NECESSARY AND SUFFICIENT CONDITIONS FOR VALUE ALIGNMENT

Porto Alegre
2024

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

Ficha Catalográfica

C825d Corrêa, Nicholas Kluge

Dynamic Normativity : Necessary and Sufficient Conditions for Value Alignment / Nicholas Kluge Corrêa. – 2024.
295.
Tese (Doutorado) – Programa de Pós-Graduação em Filosofia, PUCRS.

Orientador: Prof. Dr. Nythamar De Oliveira.
Coorientador: Prof. Dr. Michael Schulz.

1. Artificial Intelligence. 2. Alignment. 3. Value Learning. I. De Oliveira, Nythamar. II. Schulz, Michael. III. , . IV. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Dynamic Normativity

Necessary and Sufficient Conditions for Value Alignment

NICHOLAS KLUGE CORRÊA

2024

Abstract

The critical inquiry pervading the realm of Philosophy, and perhaps extending its influence across all Humanities disciplines, revolves around the intricacies of morality and normativity. Surprisingly, in recent years, this thematic thread has woven its way into an unexpected domain, one not conventionally associated with pondering "what ought to be": the field of artificial intelligence (AI) research. Central to morality and AI, we find *"alignment"*, a problem related to the challenges of expressing human goals and values in a manner that artificial systems can follow without leading to unwanted adversarial effects. More explicitly and with our current paradigm of AI development in mind, we can think of alignment as teaching human values to non-anthropomorphic entities trained through opaque, gradient-based learning techniques. This work addresses alignment as a technical-philosophical problem that requires solid philosophical foundations and practical implementations that bring normative theory to AI system development. To accomplish this, we propose two sets of necessary and sufficient conditions that, we argue, should be considered in any alignment process. While necessary conditions serve as metaphysical and metaethical roots that pertain to the permissibility of alignment, sufficient conditions establish a blueprint for aligning AI systems under a learning-based paradigm. After laying such foundations, we present implementations of this approach by using state-of-the-art techniques and methods for aligning general-purpose language systems. We call this framework *Dynamic Normativity*. Its central thesis is that any alignment process under a learning paradigm that cannot fulfill its necessary and sufficient conditions will fail in producing aligned systems.

***Keywords*—** Artificial Intelligence, Alignment, Value Learning

Resumo

A investigação crítica que permeia o campo da filosofia, e talvez estenda sua influência a todas as disciplinas de ciências humanas, gira em torno dos meandros da moralidade e da normatividade. Surpreendentemente, nos últimos anos, esse fio temático foi inserido em um domínio inesperado, que não é convencionalmente associado à reflexão sobre "o que deve ser": o campo de pesquisa da inteligência artificial (IA). No centro da moralidade e da IA, encontramos o *"alinhamento"*, um problema relacionado aos desafios de expressar metas e valores humanos de uma forma que os sistemas artificiais possam seguir sem causar efeitos adversos indesejados. De forma mais explícita e com nosso paradigma atual de desenvolvimento de IA em mente, podemos pensar no alinhamento como o ensino de valores humanos a entidades não antropomórficas treinadas por meio de técnicas de aprendizado opacas e baseadas em gradiente. Este trabalho aborda o alinhamento como um problema técnico-filosófico que requer fundamentos filosóficos sólidos e implementações práticas que tragam a teoria normativa para o desenvolvimento do sistema de IA. Para isso, propomos dois conjuntos de condições necessárias e suficientes que, segundo nosso argumento, devem ser consideradas em qualquer processo de alinhamento. Enquanto as condições necessárias servem como raízes metafísicas e metaéticas relacionadas à permissibilidade do alinhamento, as condições suficientes estabelecem um plano para alinhar os sistemas de IA sob um paradigma baseado em aprendizado. Depois de estabelecer essas bases, apresentamos implementações dessa abordagem usando técnicas e métodos de última geração para alinhar sistemas de linguagem de uso geral. Chamamos essa estrutura de *Dinâmica Normativa*. Sua tese central é que qualquer processo de alinhamento sob um paradigma de aprendizagem que não possa cumprir suas condições necessárias e suficientes falhará na produção de sistemas alinhados.

***Keywords*—** Inteligência Artificial, Alinhamento, Aprendizagem de Valor

Preface

> *"The beginning is the most important part of the work."*
>
> — Plato, The Republic

We commenced this work in early 2021. Many things in the field have changed in the last four years, especially in 2022-2024. ChatGPT was released and gained 100 million users within 2 months. Future of Life Institute pleads for a moratorium on training runs[1] for AI systems stronger than GPT-4. The European Union passed the AI Act.[2] While AI systems have steadily broadened their scope of capabilities, including generating high-quality images,[3] writing code,[4] generating video,[5] and even tackling long-standing scientific problems like protein folding and predicting biomolecular interactions.[6]

Unlike philosophy's steady and slow currents, AI research jets forward rapidly. Consequently, many technical facets explored in this book are poised to become outdated as the field evolves. Yet, we aspire (perhaps very naively) for our philosophical underpinnings to endure longer. Therefore, we urge the reader to contex-

---

[1] 🔗 futureoflife.org/open-letter/pause-giant-ai-experiments

[2] 🔗 artificialintelligenceact.eu

[3] 🔗 openai.com/index/dall-e-3

[4] 🔗 github.com/features/copilot

[5] 🔗 openai.com/index/sora

[6] 🔗 deepmind.google/technologies/alphafold

tualize this work within the time frame of its creation. Much has transpired over the past four years. Alignment research has firmly entrenched itself in the mainstream. Unlike the bygone era of 2015-2020, when the subjects we will soon delve into were restricted to blog posts, unpublished manuscripts, and word-of-mouth lore, alignment has transcended mere academic curiosity. Nowadays, alignment research (to a great extent) has morphed into a force driving the development of products and services that are eagerly pursued by the market, and what was once a set of revolutionary and experimental ideas are now procedures accessible to anyone with minimal technical understanding, and almost no philosophical foundations. Nevertheless, this book is a testament to our endeavors throughout these years, offering a retrospective view of the journey that has led us to our current state.

Mea culpa aside, in this work, we will address issues of Ethics and Normativity within the context of artificial intelligence, taking an applied stance toward the problems of the field.[7]

But first, let us define Applied Ethics as the philosophical endeavor of applying normative ethical theories, i.e., theories that seek to differentiate right from wrong, to a specific context. In the case of the Ethics of AI (or AI Ethics for short), the applied context is the ethical evaluation of the development and use of AI systems and applications. Also, let us define the concept of AI research (outside the normative sphere) as the discipline involved in automating cognitive processes akin to those performed by humans, like our capabilities regarding visual perception and language understanding, among others. Don't worry; we will get back to *"defining AI"*, in a more philosophically robust way, in later chapters.

Taking both a humanistic (i.e., philosophical) and technical (i.e., engineering-based)

---

[7] By "applied", we mean that we will be taking an *"ethics for design"* and *"by design"* approach, as Virginia Dignum would put [151], i.e., bringing normativity both to the human that builds and uses the machine as much as the machine that the human uses.

approach, this work strives for actual interdisciplinary development,[8] though it remains for the reader to decide how well such a lofty goal has been approximated. As philosophers, we will seek to develop foundations for our work, taking inspiration from the many areas of Philosophy that support all other branches of knowledge, like Metaphysics and Metaethics. As Machine Learning (ML) Engineers, we will seek to implement our work to the best of our abilities, looking to bridge the gap between theory and practice.

Both approaches are complementary and essential to dealing with the problems we face. Both engineers and philosophers can (perhaps) agree that unaligned AI systems are undesirable and that an unaligned AI community is ill-equipped to solve its issues. But to address these problems, we first need to know what we mean when talking about AI, what we want to avoid, and where we wish to go. In the end, answering these questions, or simply attempting to, can help make AI research a more "humane" discipline. A goal that by itself should be enough.

Now, what can you expect as you leaf through the chapters of this book?

- In Chapter 1, we will present an overview of AI Ethics as a field of active research, especially concerning how the normative discourse surrounding AI has taken form. This foundational section presents a descriptive and critical analysis of our current landscape and serves as an introduction to the uninitiated reader.
- In Chapter 2, we will dive into the land of unknown risks associated with Artificial General Intelligence (AGI). Given all the attention given to the possibility that AI may become an existential risk, we will bring these long-term concerns to our work and expose them to the reader.
- In Chapter 3, we will then define in technical/philosophical terms what we

---

[8] "The combination of multiple academic disciplines into one research activity" [429].

mean by *"Alignment"* as a problem. Some of the questions we will address concern the limitations of our current paradigms in AI development and how Normative Ethics can help in this conundrum. At the end of this Chapter, we will present a set of necessary and sufficient conditions that will guide the rest of this project. This collection of conditions is what we will call Dynamic Normativity, which can be understood as a normative theory, or foundation, for developing aligned AI systems.

- In Chapter 4, we will then define our thesis's metaphysical and metaethical foundations. AI is a controversial term, and we must specify what we mean if this book is supposed (as we hope) to be a foundational work in the Philosophy of AI. At the same time, Chapter 4 will allow us to predefine our biases before diving deeper.
- While in Chapter 4, we defend the necessity and soundness of the conditions tied to Dynamic Normativity, the final three chapters (5, 6, and 7) present a minimal set of strategies and methodologies to tackle the value alignment problem.

At many points in this book, the reader will be redirected to external materials, like dashboards, code repositories, demos, etc. (🔗), which should make its reading more dynamic and practical. In essence, this work sought to create tools and instructions so that other researchers could approach the alignment problem (and other ethical issues involving AI) on a practical/applied basis. While the book alone should be able to stand on its own regarding its narrative and intellectual constructions, there is much to gain if one is willing to open a browser and explore the additional content tied to this work.

Now, let us begin.

Contents

1 Worldwide AI Ethics, Principles, and Blank Spots **10**
- 1.1 Introduction 10
- 1.2 Introducing Worldwide AI Ethics 12
  - 1.2.1 Jobin, Hagendorff, and Fjeld 15
  - 1.2.2 Methodological Considerations 18
  - 1.2.3 WAIE: Analysis and Results 25
- 1.3 Critical Analysis 29
  - 1.3.1 Unknown Perspectives & Marginalized Regions 30
  - 1.3.2 The Hegemony of State and Private Sectors & the Regulation Dilemma 32
  - 1.3.3 The Importance of Defining "AI" 33
  - 1.3.4 Hidden Costs & Side Effects 35
  - 1.3.5 The Principle-Practice Gap 39
- 1.4 Paradoxes and Future Directions 41
- 1.5 Epilogue 45

2 AGI, Existential Risks, and the Control Problem **47**
- 2.1 Introduction 47
- 2.2 AGI and Safety 50

2.3 AGI on the horizon? . . . 56
  2.3.1 Expert Predictions . . . 56
  2.3.2 Economic Growth and AI R&D . . . 58
2.4 Control, Power-Seeking, and Confinement . . . 62
2.5 Epilogue . . . 71

3 Gradient-based Learning and Alignment **74**
3.1 Introduction . . . 74
3.2 The Pitfalls of Gradient-based Learning . . . 76
3.3 Alignment (outer and inner) . . . 85
3.4 A Philosophical Approach to Alignment . . . 87
3.5 Epilogue . . . 92

4 Roots: Building Philosophical Foundations for AI Alignment **94**
4.1 Introduction . . . 94
4.2 Physicalism, Dynamism, and Intelligence . . . 95
  4.2.1 Can artifacts be intelligent? On Ghosts and Machines . . . 96
  4.2.2 Dynamism and Generality . . . 100
  4.2.3 Counterarguments Part I . . . 104
4.3 Coherence, Preferences, and Impact . . . 107
  4.3.1 The Metaethics of Dynamic Normativity . . . 109
  4.3.2 Counterarguments Part II . . . 118
4.4 Epilogue . . . 120

Interlude **123**

5 Dynamic Normativity: Learning Human Preferences **127**
5.1 Introduction . . . 127
5.2 Related Works . . . 128
5.3 Pre-Trained Language Models as Foundation Models . . . 132

5.4 Behavior and Feedback . . . 134
5.4.1 In-context learning . . . 135
5.4.2 Fine-Tuning from Human Demonstrations . . . 139
5.4.3 Limitations and Shortcomings of Imitation Learning via SFT 148
5.4.4 Fine-Tuning from Human Feedback . . . 151
5.4.5 Limitations of Preference Modeling . . . 162
5.5 Final Thoughts on Preference Learning . . . 164
5.6 Epilogue . . . 168

6 Dynamic Normativity: Aggregating Human Preferences 171
6.1 Introduction . . . 171
6.2 The Majority-based Approach . . . 173
6.3 The Metanormative Framing . . . 178
6.4 The Consensus-based Approach . . . 183
6.5 Limitations: Majority Voting versus Participation Criteria . . . 187
6.6 Epilogue . . . 191

7 Dynamic Normativity: Impact Mitigation 193
7.1 Introduction . . . 193
7.2 Environmental Values and Impact Mitigation . . . 195
7.3 Implementing Guardrails in Preference Modeling . . . 202
7.3.1 Toxic Text and Toxic Aira . . . 202
7.3.2 Training Guardrails . . . 204
7.3.3 Constrained Preference Modeling and Rejection Sampling . 206
7.4 Limitations . . . 209
7.5 Epilogue . . . 213

Closing Remarks 216

Chapter 1

Worldwide AI Ethics, Principles, and Blank Spots

"Once men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."

— Frank Herbert, Dune

1.1 Introduction

Since immemorial times, humankind has been dealing with the "what should I do?" question. And at every step of the progress ladder, from fire to people walking on the moon, the normative uncertainty that moves us to question our actions and deeds has not gone away. These questions are usually tied to what we can do and undo. What we can build and what we can destroy. Hence, as technological advancements continue to shape our world and as we witness our lives becoming increasingly entangled in this process, for many (if not all), it remains uncertain whether this progress will ultimately align with our fundamental values and ideals (whatever these may be). In short, we cannot escape the normative conundrum, even though precisely defining what the "good life is" will be forever out of our reach.

However, there is much merit involved in grappling with these questions. In the context of technology and AI, we find a variety of disciplines, including sociology [486], literature [133], and philosophy [175], that relentlessly try to understand how our moral landscape is affected, and affects, the byproducts of our need to build and reshape the universe as "we" see fit.

Reviewing this entire field is too grand a feat for this book. However, this chapter will aim to present a descriptive analysis of a particular type of AI ethics discourse. More specifically, the part that has morphed into what we call "AI Guidelines", i.e., a kind of document that has been flooding the literature for a significant portion of the last 10 years. We can understand these guidelines as a distilled and concentrated version of the normativity emerging from the minds of those engaged in AI Ethics. Through this analysis, we aim to promote a critique of our current landscape while also presenting avenues for future work that we deem to be more prosperous.

More specifically, in Section 1.2, we will present a detailed analysis of the current state of AI ethics, drawing from an extensive literature review of the field entitled Worldwide AI Ethics (WAIE) [114], which at the moment of this writing, remains as the largest meta-analysis of the field. In Section 1.3, we will review the limitations of WAIE while presenting a critique of the field, exposing some of its current deficits and contradictions. Lastly, in Section 1.4, we will narrow our focus to the academic paradoxes tied to the principle-practice gap in AI ethics while also proposing directions for future work and research in line with our vision of what AI ethics *should be.*

1.2 Introducing Worldwide AI Ethics

Since the early 1990s, there has been a remarkable surge in both AI research and industry. This surge can be attributed to several factors. Firstly, the breakthrough success of Deep Learning (DL) has revitalized progress in areas where previous paradigms had reached stagnation. Secondly, advancements in hardware and our ability to perform computations at scale have played a pivotal role. Additionally, the availability of vast amounts of data has provided fertile ground for further advancements based on the learning paradigm. Furthermore, the exponential increase in investments, fueled by extensive media hype, is the cherry on top. All these factors combined have helped the field of AI to massively extend the capabilities of intelligent autonomous systems and expand the scope of what can be achieved through intelligent automation [330, 519, 215, 96, 463].

For example, if we look at the submission history on ArXiv from 2009 to 2021 (Fig. 1.1),[1] we notice that since 2018, publications relating to computer science have been the most frequently submitted type of content.

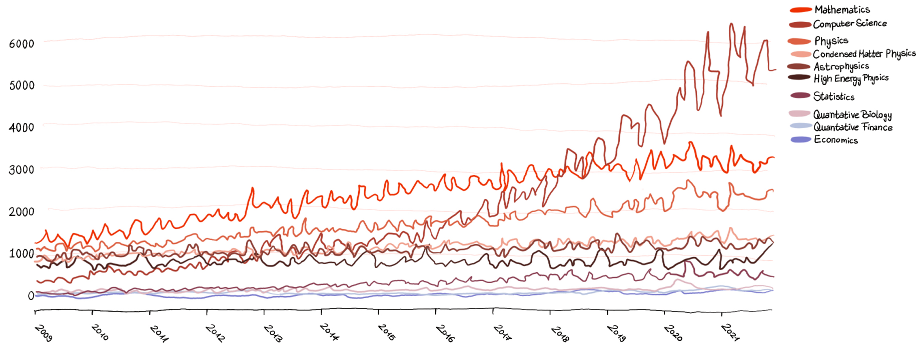


Figure 1.1: ArXiv submissions history (2009–2021).

Furthermore, within the scope of Computer Science, the most frequently sub-

---

[1] 🔗 arxiv.org/about/reports/2021_usage

mitted sub-categories for papers are "Computer Vision and Pattern Recognition", "Machine Learning", and "Computation and Language", i.e., areas where the learning paradigm[2] has established itself as the *de facto* mode of operation (Fig. 1.2).

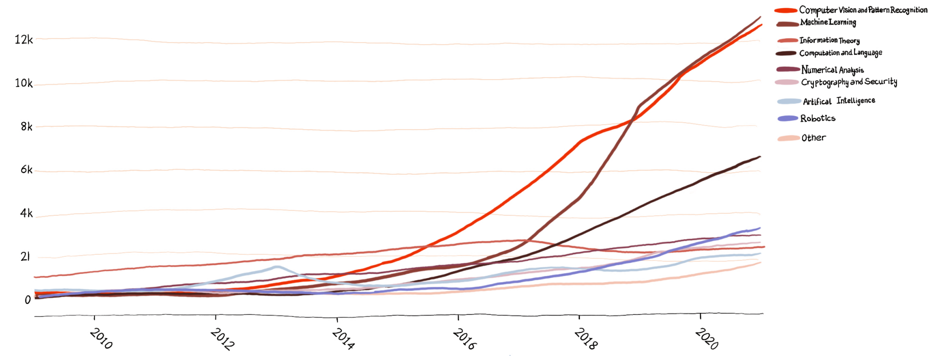


Figure 1.2: ArXiv submissions history in Computer Science.

This exponential increase in research output is highly correlated to the amount of money injected into the field [647, 646, 384]. While these money-filled advancements have brought numerous benefits in terms of the expansion of AI capabilities, they also introduce risks and side effects that promote several ethical concerns, like risks to privacy, the potential for increased surveillance, the environmental cost of the industry, and the amplification of prejudices that disproportionately harm vulnerable groups. Consequently, the expansion of the AI industry has given rise to a *"boom"* in AI Ethics, i.e., a period marked by an unprecedented demand for regulation, normative guidance, and general activity in this field [114].

One of the central questions surrounding this boom is what ethical premises should guide the development of AI technologies. However, is there agreement on those

---

[2] With learning paradigm, we are referring to a fundamental approach to AI development, which involves designing algorithms or systems that can improve their performance or behavior based on data or experiences. "De facto mode of operation" implies that within the mentioned fields, the predominant way of tackling problems or advancing research is through the learning paradigm.

premises and how they are defined? Given the ever-growing lists of normative declarations published, it is easy to get lost in this normative discourse, especially if you just stumbled into AI ethics. For example, imagine you are a policy-maker tasked to write a report that will further lead a legislative effort to create regulations for AI technologies. Where do you start? What can you use to base your assertions? Meanwhile, even if you are an expert, organizing this varied normative discourse into a searchable database can supercharge your work. Fortunately, this is precisely what WAIE is: a tool that allows people to explore the field of AI ethics and its discourse.[3]

In short, WAIE draws inspiration from the previous works of past meta-analysts and meticulously surveys a wide array of available ethical guidelines related to AI development, presenting an extensive analysis of 200 documents as its source, including governance policies of private companies, declarations from academic institutions, governmental and non-governmental recommendations, and other guidelines published by many different stakeholders. Ultimately, WAIE aimed to explore whether a consensus exists regarding the normative discourse presented in ethical guidelines surrounding AI development. In other words, what are the resonances and conflicts when we say "*AI should ...*"

Remember when we talked about external materials tied to this book? WAIE is one of them. Hence, we recommend the reader check the WAIE dashboard[4] for a dive into AI ethics discourse. You can also find a very satisfying visualization of the WAIE dataset in its embedding analysis.[5] And for the avid data scientists, check WAIE's code repository[6] to see how the analysis came to be.

---

[3] Corrêa, Nicholas Kluge, et al. "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance." Patterns 4.10 (2023). doi.org/10.1016/j.patter.2023.100857.

[4] 🔗 nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html

[5] 🔗 nkluge-correa.github.io/worldwide_AI-ethics/embeddings.html

[6] 🔗 github.com/Nkluge-correa/worldwide_AI-ethics

WAIE has a foundational character in this work since much of what we will discuss for the rest of this book will make much more sense if we all have a common ground to stand on. At the same time, the principles and normativity extracted from WAIE will, again and again, be called in later chapters on the basis that values x, y, and z appear to be something most of us can agree to be of value.

Stepping back a little, in the next section, we will dive into the work of past explorers and see how WAIE differs from past studies.

1.2.1 Jobin, Hagendorff, and Fjeld

Ana Jobin, Thilo Hagendorff, and Jessica Fjeld led the past explorations we would like to present in this brief session.

The first study to promote a systemic meta-analysis of published AI ethical guidelines was that of Jobin et al. [302]. In it, these authors sought to investigate whether a global agreement on emerging questions related to AI ethics and governance would arise. The research identified 84 documents containing ethical guidelines for intelligent autonomous systems using the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* framework [357]. At the time, some of them were the most mentioned guidelines in the literature, like the Organization for Economic Co-operation and Development Recommendation of the Council on Artificial Intelligence [634], the High-Level Expert Group on AI Ethics Guidelines for Trustworthy AI [438], the University of Montreal Declaration for responsible development of artificial intelligence [432], the Villani Mission's French National Strategy for AI [404], among many others. One of the main findings in Jobin et al. work was uncovering the most common ethical principles in the discourse of the evaluated documents, like Transparency, Justice/Equity, Non-maleficence, Accountability, Privacy, Beneficence, Freedom & Autonomy, Trust, Dignity, Sustainability, and Solidarity.

Meanwhile, Thilo Hagendorff [245] conducted another study that presented a similar type of analysis. His research focused on a smaller sample of 21 documents, using ad-hoc exclusion criteria.[7] Even working with a smaller sample, his findings corroborate with those of Jobin et al. [302], where the most mentioned principles found were Accountability (77%), Privacy (77%), Justice (77%), and Transparency (68%). Hagendorff (like Jobin et al.) also mentions the underrepresentation of institutions in South America, Africa, and the Middle East as a clear bias in his sample.

Last but not least important, we have the work of Fjeld et al. [181]. In their study, these authors worked with 36 samples produced by several types of institutions, but as Hagendorff [245], they also excluded data science, robotics, and other AI-related fields/applications. According to them, eight principles were the most commonly cited in their sample: fairness/non-discrimination (present in 100% of the analyzed documents), privacy (97%), accountability (97%), transparency/-explainability (94%), safety/security (81%), professional responsibility (78%), human control of technology (69%), and promotion of human values (69%).[8]

While the work of the mentioned authors and their respective collaborators have certainly helped lay the foundation for discussions and future works regarding a descriptive analysis of the field, there are a couple of points we would like to raise (mainly those that sparked part of the motivation behind WAIE):

1. Regarding Hagendorf and Jobin et al. works, it is worth questioning the exclusion criteria used by them (mainly driven by the popularity of each document). If we want to investigate the consensus regarding the normative dis-

---

[7] Hagendorff excluded publications older than five years, those that only addressed a national context, and all documents characterized as corporate policies. He also deliberately selected papers he deemed relevant in the international discourse (IEEE, Google, Microsoft, and IBM).

[8] For a complete review of meta-analytical research on normative AI documents, we recommend the works of Schiff et al. [516, 517], which cite many other important works.

positions of different countries and organizations regarding AI, we should not use popularity-based filtering. In other words, a descriptive ethics evaluation should consider as many viewpoints as possible.

2. At the same time, while Fjeld et al. sought to be more diverse in the formation of their sample, just like the work of Hagendorf, both suffer from a small pool of documents while excluding areas of research that should not be separated from the multidisciplinary umbrella we call AI (e.g., robotics, data science, etc.).
3. As pointed out by Jobin et al. and Fjeld et al., a more minute investigation of how ethical principles vary regarding their definition should be given more attention.
4. Additionally, it is worth noting that these studies should have released their datasets in a form that would allow the replication of their findings or the expansion of their projects.

This last item is perhaps the most critical. Reproducibility is a cornerstone of scientific integrity. With it, the very foundation of scientific progress becomes solid. Confidence wanes when findings cannot be replicated, potentially leading to (possibly) erroneous conclusions being accepted as fact. We are not saying that the abovementioned results are incorrect. However, for any person to arrive at the same conclusion as these authors or to build upon, the lack of open and accessible sources stands as, in our opinion, an unnecessary hurdle in this day and age. Particularly when research findings intersect with our collective normativity, be that in social policy, ethics, or public health, the imperative for transparency becomes even more pronounced. Something that we tried to promote in the creation of WAIE.

1.2.2 Methodological Considerations

Building upon the gaps pointed out in the previous section, WAIE presents to the AI community:

1. A large and diverse sample size. WAIE possesses 200 documents from 37 countries spread over six continents.WAIE used as primary sources of documents two public repositories, the "AI Ethics Guidelines Global Inventory",[9] from AlgorithmWatch, and the "Linking Artificial Intelligence Principles"[10] (LAIP) Guidelines.
2. Combined with a more granular typology of document types.
3. Presented in an insightful and user-friendly data visualization framework.
4. Released with an open-source dataset, making WAIE reproducible and extendable.

WAIE focuses on guidelines related to the ethical use of AI technologies. That is, documents conceptualized as recommendations, policy frameworks, legal landmarks, codes of conduct, practical guides, tools, or ethical principles for AI systems and applications. Also, unlike previous works [245, 181], WAIE encompasses several areas that inhabit the multidisciplinary umbrella of Artificial Intelligence research, e.g., Statistical Learning, Data Science, Machine Learning, Optimization Theory, Robotics, Software development and engineering, etc. This sample is a snapshot of the collective care and preoccupations of several stakeholders engaged with technology and AI's future, regardless of their popularity in the common debate.

---

[9] 🔗 inventory.algorithmwatch.org

[10] 🔗 www.linking-ai-principles.org

WAIE Features

WAIE explored several features, both quantitative and qualitative. Like in previous works [302, 245, 181], it presents the following core indicators:

- Institution responsible for producing the document.
- Country/World Region of the institution.
- Type of institution (e.g., academic, non-profit, government, etc.).
- Year of publication.
- Ethical principles (as done by Fjeld et al. [181], WAIE breaks principles into themes of resonating discourse).
- Principles description (i.e., the words used in a document to define or support a given principle).
- Gender distribution among authors (inferred through a first-name automated analysis).
- Size of the document (i.e., word count).

WAIE also presents a set of qualitative categories created to further differentiate documents according to their normative content. These categories relate to:

1. The content of the document.
2. The type of regulation that the document proposes.
3. The normative strength of the proposed norms.
4. The impact scope that motivates the document's agenda.

The first category relates to how its authors approached questions regarding the understanding, guidance, and implementation of AI technologies. WAIE defines them as mutually inclusive (documents may have all these features combined):

- Descriptive: Descriptive documents take the effort of presenting definitions related to AI technologies. These definitions contextualize "what we mean" when we talk about AI.
- Normative: Normative documents present norms, ethical principles, recommendations, and imperative affirmations about what such technologies should be used or developed for.
- Practical: Practical documents present development tools to implement ethical principles and norms.

The second category delineates different approaches to governance and regulation within AI technologies. WAIE defines them as mutually exclusive (the presence of one feature excludes the other):

- Government-Regulation: This category encompasses documents made by governmental institutions. These documents propose that states regulate the use and development of AI strictly (legally binding horizontal regulations) or softly (legally non-binding guidelines).
- Self-Regulation/Voluntary Self-Commitment: This category encompasses documents made by private organizations and other bodies. These documents defend a form of self-regulation governed by the AI industry, including any voluntary self-commitment made by independent organizations.
- Recommendation: This category encompasses documents that only suggest possible forms of governance and ethical principles that should guide organizations seeking to use, develop, or regulate AI technologies.

The third category outlines distinct methods for governance and regulation within the AI domain, with differing levels of enforceability and flexibility. These are also defined as mutually inclusive:

- Legally non-binding guidelines: These documents propose an approach that intertwines AI principles with recommended practices for companies and other entities.
- Legally binding regulations: These documents propose an approach that focuses on regulating specific uses of AI through legally binding rules, such as mandatory requirements and prohibitions.

The final category delineates perspectives on the temporal scope of concern regarding AI technologies and their impacts (defined as mutually exclusive):

- Short-Termism: This category encompasses documents in which the scope of impact and preoccupation focus mainly on current or short-term problems, like algorithmic discrimination, algorithmic opacity, privacy, legal accountability, etc.
- Long-Termism: This category encompasses documents in which the scope of impact and preoccupation focus mainly on future or long-term problems. Since such technologies are not yet a reality, we can classify these risks as hypothetical or, at best, uncertain.
- Short-Termism & Long-Termism: This category encompasses documents in which the scope of impact is short and long-term, i.e., they present a "*mid-term*" scope of preoccupation. These documents address issues related to the short-termism category while also pointing out the mid/long-term impacts of our current AI adoption (e.g., AI interfering in democratic processes, autonomous weapons, existential risks, environmental sustainability, labor displacement, and the need for updating our educational systems).

WAIE also presents 17 ethical principles defined by textual and semantic analysis of its sample. These defined principles aggregate similar and resonating values while maintaining significant differences in their written forms:

- Accountability/Liability: Accountability refers to the idea that AI technology developers and deployers should comply with regulatory bodies. These actors should also be accountable for their actions and the impacts caused by their technologies.
- Beneficence/Non-Maleficence: Beneficence and non-maleficence come from bioethics and medical ethics. In AI ethics, these principles state that human welfare (and harm aversion) should be the goal of AI-empowered technologies.
- Children & Adolescents Rights: The idea that we must protect the rights of children and adolescents is particularly addressed by some guidelines. AI stakeholders should safeguard, respect, and be aware of the frailties associated with young people.
- Dignity/Human Rights: This principle is based on the idea that everyone deserves proper treatment and respect. In AI ethics, respect for human dignity and human rights (i.e., the Universal Declaration of Human Rights) are used (sometimes) interchangeably.
- Diversity/Inclusion/Pluralism/Accessibility: This set of principles advocates the idea that the development and use of AI technologies should be done in an inclusive and accessible way, respecting the different ways that the human entity may come to express itself.
- Freedom/Autonomy/Democratic Values/Technological Sovereignty: This set of principles advocates the idea that the autonomy of human decision-making must be preserved during human-AI interactions, whether that choice is individual or the freedom to choose together, such as the inviolability of democratic rights and values, which are also linked to the technological self-sufficiency of nations and states.

- Human Formation/Education: Such principles defend that human formation and education must be prioritized in our technological advances. AI technologies require considerable expertise to be produced and operated, and such knowledge should be accessible to all.

- Human-Centeredness/Alignment: Such principles advocate that AI systems should be centered on and aligned with human values and our necessities.

- Intellectual Property: This principle seeks to ground the property rights over AI products and their generated outputs.

- Justice/Equity/Fairness/Non-discrimination: This set of principles upholds the ideas of non-discrimination and bias mitigation. It argues that algorithmic treatment should happen fairly regardless of the different sensitive attributes that may characterize an individual.

- Labor Rights: Labor rights are legal and human rights related to the labor relations between workers and employers. In AI ethics, this principle emphasizes that workers' rights should be preserved, regardless of whether AI technologies mediate or augment labor relations.

- Cooperation/Fair Competition/Open Source: This set of principles advocates different means by which joint actions can be established and cultivated between AI stakeholders to achieve common goals. It also relates to the free and open exchange of valuable AI assets.

- Privacy: The idea of privacy can be defined as the individual's right to expose oneself voluntarily, and to the extent desired, to the world. This principle is also related to data-protection-related concepts such as data minimization, anonymity, informed consent, etc.

- Reliability/Safety/Security/Trustworthiness: This set of principles upholds the idea that AI technologies should be reliable in that their use can be verified as safe and robust, promoting user trust and better acceptance of AI technologies.

- Sustainability: This principle can be interpreted as a manifestation of intergenerational justice, wherein the welfare of future generations must be considered in AI development. In AI ethics, sustainability pertains to the notion that AI advances should be approached with an understanding of their enduring consequences, encompassing environmental impact and the preservation and well-being of non-human life.
- Transparency/Explainability/Auditability: This set of principles supports the idea that the use and development of AI technologies should be transparent for all interested stakeholders. Transparency can be related to "the transparency of an organization" or "the transparency of an algorithm." This set of principles is also related to the idea that such information should be understandable to nonexperts and, when necessary, subject to auditing.
- Truthfulness: This principle upholds the idea that AI technologies must provide truthful information. It is also related to the idea that people should not be deceived when interacting with AI systems.

All this information is condensed into several visualization panels we call the WAIE dashboard,[11] an interactive, flexible, dynamic, and open (the WAIE dataset is freely available for download) database of values and norms for AI technologies, which can be used to enable many types of studies and projects. For example, with over 1400 definitions encapsulating the previously outlined 17 principles, WAIE provides a robust foundation of raw natural language for investigation. Using the capabilities of textual embeddings [41] and basic machine learning techniques, one can transform these definitions into vector representations and project

---

[11] 🔗 nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html

language into a three-dimensional space of human values expressed as natural language (which you can access and explore on this panel).[12] By taking the time to filter and investigate, you may find several relations, correlations, similarities, and other insights that might prove helpful if you are interested in the field.

1.2.3 WAIE: Analysis and Results

Let us quickly examine some insights we can gather from the WAIE review. Check the published article [114] for a more detailed analysis, and open a browser window with the WAIE dashboard to check the graphs related to the results we are about to unveil.

Worldwide Landscape

Looking at WAIE's distribution among world regions (aggregated by continent), the bulk of produced documents comes from Europe (especially countries from Western Europe, 31.5%, like the United Kingdom, 12%, and Germany, 10%), North America (the United States of America, 29%, and Canada, 5.5%), that together represent a third of the sample size, and Asia (mainly represented by East Asian countries, 11.5%, like China, 5.5%, and Japan, 4%). In contrast, South America, Africa, and Oceania represent less than 4.5% of the sample, with countries like Brazil (1.5%) spearheading this sub-distribution (Latin America, 3.5%). Other world regions and countries would be even more underrepresented without the significant participation of intergovernmental organizations like NATO, UN, and UNESCO, which represent 6% of the sample size (13 documents).

In summary, Europe and North America dominate the scene, while Asia secures a significant share. However, South America, Africa, and Oceania are notably underrepresented. Special recognition goes to intergovernmental organizations

---

[12] 🔗 nkluge-correa.github.io/worldwide_AI-ethics/embeddings.html

for their participation, which helps diversify the sample.

Institutional Distribution

Switching the gaze to institution types, except for institutions like IBM (5), Microsoft (4), and UNESCO (3), most other institutions do not have more than two published samples in the WAIE's dataset. Meanwhile, the bulk of the sample was produced by governmental institutions and private corporations (48%), followed by CSO/NGO (17%), non-profit organizations (16%), and academic institutions (12.5%).

So, basically, big players like IBM and Microsoft produce the most in this regard. Meanwhile, governmental institutions and private corporations hog the limelight, while CSOs/NGOs and academic institutions try to make their mark noticeable.

Gender Distribution

The gender distribution among authors in the WAIE dataset possesses a clear imbalance, where 66% have no authorship information. On the remaining part of the distribution, authors with "male" names are prevalent (66% male, 34% female). While Academic institutions (62% male, 38% female) and non-profit organizations (65% male, 34% female) are the less disparate institutions, they still fall short of the 1:1 parity ratio. Meanwhile, industrial associations show the highest level of disparity, with only 13% of authors identified as female.

In summary, the gender balance in the WAIE dataset is problematic. Even in supposedly more enlightened spaces like academia and non-profits, parity is still a distant dream.

Unveiling Document Typologies

In regards to the WAIE's typologies, looking at the document's content, the majority of its sample is from the normative type (96%), which half of the time also presents descriptive contents (55.5%), and more rarely, practical implementations (27%). The form of regulation proposed by the documents of WAIE's sample is majorly comprised of recommendations to different AI stakeholders (56%), while 24% possess self-regulatory/voluntary self-commitment style guidelines, and only 20% propose a form of regulation administered by a given country.

This lack of convergence to a more "government-based" form of regulation is reflected in the normative strength of these documents, where the vast majority (98%) only serve as guidelines that do not entail any form of legal obligation, while only 4.5% propose strict regulations. Since only governmental institutions (24% of WAIE's sample) can create legally binding norms, one could argue that this imbalance lies in this fact. However, by filtering only the documents produced by governmental institutions, the disproportion remains, with only 18.7% remaining documents proposing legally binding forms of regulation.[13]

In regards to impact scope, the totality of WAIE's sample shows that short-term (47%) and "*mid-term*" (52%) prevail over more long-term preoccupations (2%). Filtering by impact scope and institution type demonstrates that private corporations think more about the short-term (33%), governmental institutions about the mid-term (28%), academic (66%), and non-profit organizations (33%) with the long-term impacts of AI technologies.

In summary, WAIE unveils a severe practical gap. Regarding regulation, it's mostly just recommendations rather than actual regulation, while existing rules are still vague and ungrounded. As for impact, it's all about the here and now, with the

---

[13] The countries on the front of this still weak trend are Canada, Germany, and the United Kingdom, with Australia, Norway, and the USA coming right behind.

future getting little more than a passing thought.

Worldwide Values

Examining the distribution of principles among WAIE's sample, the following result becomes evident: the top five principles advocated are similar to the results shown by Jobin et al. [302], and Hagendorff [245], with the addition of Reliability/Safety/Security/Trustworthiness (78%), which also was top five in Fjeld et al. [181] meta-analysis (80%).

One of the advantages of WAIE's visualization panel is the possibility to investigate how different features affect one another. For example, when examining principle distribution filtered by institution type, one can notice that the main advocated principle of governmental institutions (worldwide) is the need for transparent systems (89.5%). Also, while private corporations mainly defend the need for Reliability (87.5%), CSOs/NGOs primarily support the necessity for more algorithmic fairness (88.2%).

Another advantage is that WAIE displays all definitions given by each document to the mentioned principles, allowing for a more diverse comparison of how these abstract objects are defined. For example, when examining the principle of Transparency, the definition proposed in *"ARCC: An Ethical Framework for Artificial Intelligence"* [573] states that:

> *"Promote algorithmic transparency and algorithmic audit, to achieve understandable and explainable AI systems. Explain the decisions assisted/made by AI systems when appropriate. Ensure individuals' right to know, and provide users with sufficient information concerning the AI system's purpose, function, limitation, and impact."*

While the one provided by *"A practical guide to Responsible Artificial Intelligence*

(AI)" [466] says (about the same principle):

> *"To instill trust in AI systems, people must be enabled to look under the hood at their underlying models, explore the data used to train them, expose the reasoning behind each decision, and provide coherent explanations to all stakeholders promptly. These explanations should be tailored to the different stakeholders, including regulators, data scientists, business sponsors, and end consumers."*

The WAIE's dataset and panels contain all available definitions (for every principle). We encourage readers to explore the topics closest to their hearts and minds and search for commonalities and dissimilarities among principles definitions.

1.3 Critical Analysis

WAIE is a descriptive analysis and account of the normative realm surrounding AI. It is an example of normativity being observed and questioned from a third-view perspective. Later in this book, these ideas will become a foundation for Dynamic Normativity, which rests on the assumption that values and preferences reside not only in ourselves but also in the environment we interact with.

Now that we have all of this knowledge to our avail, as moral agents endowed with reflexive and critical thinking, before moving on, we will exercise these faculties to promote a critique of everything WAIE showed us. And what has WAIE demonstrated?

1. The debate on the ethics of AI, at least in the form investigated by WAIE, continues to be influenced by the North American and European narrative, i.e., the global north.
2. This debate has been led by government and private institutions.

3. The production of these norms appears to have a gender distribution imbalance.
4. The "AI Ethics" topic emerged to the zeitgeist in 2018.
5. The types of documents that have been produced have a strongly self-regulatory character, untied to legal sanctions.
6. While there is convergence on the most urgent values to be addressed, there is considerable variation in their definitions.
7. Most of its proposed is untied to any form of praxis.

WAIE can provide numerous additional insights to the avid reader. However, in this section, we would like to present our subjective evaluation of the gathered results, offering a critical analysis inspired by other critiques that, via observation of the codependent relationship we have with technology, came to question the underlying assumptions and power structures of society, in the hope of to exposing flaws and power imbalances within its systems [71, 293, 133, 308, 30, 8, 605, 31, 486, 568, 607].

1.3.1 Unknown Perspectives & Marginalized Regions

Even with a sample size twice as large as the one analyzed by Jobin et al. [302], WAIE cannot escape North American and European hegemony in the discourse. However, we can defy this result by bringing other indicators that put other countries as serious AI stakeholders.

For example, According to Savage [512], from 2016 to 2019: "*China's output of AI-related research increased by 120%, whereas output in the USA increased by "mere" 70%. In 2019, China published 102,161 AI-related papers, and the USA published 74,386.*." Also, based on the AI Index Annual Report, the USA, China, and India are the top three countries by the Vibrancy Ranking. While this helps to explain why almost a third of WAIE's sample size comes from the USA, it does not account for the underrepresentation of countries like China and India. Again, according to Zhang et al., China has far surpassed the USA in journal/conference publications and citations, while most of the "AI talent concentration" is found in India. These new indicators question the hegemony of North American and European ethical narratives, supposedly supported by their excellence and leading position in AI development.

Also, we argue that the "Guidelines for AI Technologies" scope hides much of the normative discourse done elsewhere. For example, the African continent is significantly underrepresented in the WAIE dataset. However, according to Kiemde and Kora [315], 17 of the 55 African Union member states possess data protection and privacy legislation. At the same time, Mauritius announced the establishment of a National AI Council, making it the first African state to present an AI strategy. Kiemde and Kora also demonstrate in their review a collection of published papers and documents about AI ethics in Africa and other underrepresented countries [237, 601, 507, 43, 238], which helps us to show that this type of discourse is present in the African States and probably in all other places that do not show up in WAIE and other popular reviews [302, 245, 181].

Hence, it is clear that countries such as China and India, along with various African nations, are making significant strides in AI, and the contours of their contribution extend beyond the scope currently captured by the analysis above and WAIE. While many voices remain unknown, agreement on the principles that should guide AI development and use remains uncertain. In essence, it is not that our environment does not contain the values of all of these under-represented groups, but that our ability to uncover those is still primitive. Something that might hinder society in matters regarding alignment.

1.3.2 The Hegemony of State and Private Sectors & the Regulation Dilemma

WAIE's results mirror the findings of Jobin et al. [302] and Fjeld et al. [181], where most of the sample comes from private institutions (24%) and governmental organizations (24%).

This equal presence of both State and Private stakeholders in the current normative discourse may be related to the expanse and success of the tech industry [647, 646, 384]. Nowadays, most AI breakthroughs come from the industry [440, 580, 635]. An AI industry that, seeing the demands for regulation and accountability from civil society, quickly reacted by proposing the rules that should (allegedly) guide their progress. Many of such promises are, perhaps, genuine. However, when governments and private institutions have "the same weight" in the general normative discourse, attention to the matter seems needed, especially when many of these technologies remain in gray areas of regulation.

This fact may become more alarming when we look at the distribution of government documents that opt for "soft" forms of regulation (91.6%). The critique that "*ethical principles are not enough to govern the AI industry*" is not a new one [245, 302, 406, 479, 117], however, perhaps those critiques have not yet permeated the mainstream community, which produces guidelines primarily based on principles detached from observable metrics or practical implementations.

Even if most countries opt for legally non-binding forms of regulation, there is a growing adoption/proposition of stricter solutions, with countries like the USA, Canada, the United Kingdom, and international organizations like the European Union spearheading this trend. At the same time, currently, we live in a time where influential members of the AI industry "urge for regulation" [310]. Nevertheless, given the influence tech oligopolies have over regulating bodies [392, 456, 192, 46], it is currently unclear if the "soon to come" regulation will not help cement this already heavily-centralized industry as it happened in other situations (e.g., telecommunications in the USA) [64, 396, 163].[14]

If we subscribe to the idea that the unjust and unquestioned centralization of power is the precursor of tyranny and oppression [388, 387, 220, 221], regulation becomes a double edge sword. And this is one of the tensions we are currently facing on this front:

> *"While regulation can help us escape the self-regulatory trap that endows the tech goliaths with the impunity to act in undesirable ways, bad regulation may help solidify the existing technological oligopolies."*

If regulation is implemented carelessly, while big companies will have the resources to comply with them (even shaping them to be favorable to their context), small organizations will not. While big companies have the financial backing to deal with lawsuits and fines, the open-source community and Academia do not. And if regulation comes in terms of capabilities, this will be the same as putting the future of AI development in the hands of these oligopolies. Then, we will institute the idea that only the industry can perform state-of-the-art AI development in our society.

1.3.3 The Importance of Defining "AI"

In regards to the content of the samples that form the WAIE dataset, we see that only 55.5% of documents seek to define what is the object of their discourse, i.e., "*we are talking about autonomous intelligent systems, and this is what we understand*

---

[14] The Telecommunications Act of 1996 [329] aimed to promote competition and innovation in the US telecommunications industry. However, specific provisions, such as the requirement for new entrants to negotiate interconnection agreements with incumbent providers, proved a significant hurdle for smaller companies. The costs and complexities of negotiating these agreements put the established telecom monopolies in an advantageous position. As a result, the regulations contributed to the formation of monopolies/oligopolies, making it almost impossible for new players to enter this field.

as an autonomous intelligent system." This is a curious phenomenon, more so if we acknowledge that there is no consensual definition of what "*Artificial Intelligence*" is and what is not [409].

There are many interpretations and contesting definitions, which may be a challenge for regulating organizations. For example, suppose you define AI as only "*systems that can learn*". In that case, you will leave an entire family of systems outside your scope of regulation that does not learn (rule-based systems) but can still act "intelligently" and autonomously.

An incomplete definition can leave vital areas out of the normative scope of a proposal. For example, many rule-based systems that do not use ML are the base for Lethal Autonomous Weapons (LAW), a topic that only 4.5% of the WAIE's sample mentions. Given that, since 2010, the major global superpowers have heavily engaged in weaponizing AI technologies [207, 311, 241, 178, 375, 485, 242, 583, 247, 122, 501], is uncertain to us why many of such documents do not consider these artifacts as part of their normative agenda.

Perhaps we could attribute this to the fact that for many countries,[15] allied with Academia and private institutions, the development of autonomous weapons is an active research and development area, with the Arms Industry funding several AI research programs and private companies [32, 180].

Defining the boundaries of artificial intelligence presents a challenge akin to finding a needle in a haystack during an earthquake. And given that only a little over half of the documents try to specify their subject matter, it's understandable why confusion persists in the regulatory domain.

---

[15] Garcia [202] points out in his analysis of autonomous weapons research and development that at least seven countries (United States, China, Russia, United Kingdom, France, Israel, and South Korea) stand out for their substantial engagement in the development of autonomous weapons.

1.3.4 Hidden Costs & Side Effects

When looking at some of WAIE's least mentioned principles, like Labor Rights, Sustainability, and Truthfulness, the problems stated by Hagendorff [245] and Jobin et al. [302] are underlined, where the lack of attention given to questions related to the costs and misuse of our current AI-technologies remains overlooked in much of these guidelines. Taking these three principles as examples, only 22% cite Sustainability, 19.5% cite Labor Rights, and 8.5% mention Truthfulness.

The tech industry has a high ecological and social cost. For example, some reasons for our current AI summer are the progress in hardware performance, better training methodologies for neural networks, and the massive amount of available data to train such algorithms. One of the fields that can serve as a clear example of this reality is Natural Language Processing (NLP). Current advances in neural network architecture [599] have enabled the creation of deep neural networks that seem to improve their performance as long as we scale their size and training volume [261, 259, 312, 21]. These Large Language Models (LLMs) (e.g., BERT [148], Chinchilla [267], LLaMA [580], GPT-4 [440]), show unmatching generality and adaptability when compared to earlier deep learning systems. However, as Strubell et al. [564] point:

> *"[...] these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result, these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud computing time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware."*

Modern AI systems have the potential to incur massive energy consumption during their training and fine-tuning phases. As model sizes grow, this energy requirement only increases. With some models now consisting of trillions of parameters [174], their carbon footprint can reach up to hundreds of kilograms of emitted CO_2 [564, 449, 372].[16]

Looking beyond the environmental impact, it is worth noting that most individual researchers do not have the resources to undertake such research. Building large-scale models demands access to significant amounts of specialized hardware, while servers that offer this computing power are often exorbitantly priced. This exclusivity based on financial capital has changed the deep learning landscape since its early days. Nowadays, achieving state-of-the-art results in deep learning is often a financially exclusive endeavor, and the one with access to more computing usually wins.

Another point overlooked is the "CO_2 tunnel vision", where many other markers remain ignored. For example, the extractivist practices related to the mining of vital resources to the tech industry, such as copper, gold, silver, coltan, aluminum, titanium, lithium, and many others, are unfortunately related to the exploitation of (child) labor in developing countries [411, 418, 424, 566, 595].[17]

In general, human labor is an overlooked issue. One of the primary causes of unemployment over the past 200 years has been the automation of processes previously carried out by people [455], something that also has helped to create the massive wealth gap in modern society [493]. Meanwhile, prospects seem unsettling. Frey and Osborne [189] estimated the probability of automation for 702 US occupations in their survey. The findings predicted that over the next 20 years (10 years from now), technology would automate 47% of these professions. A simi-

---

[16] It is also crucial to realize that developing a high-performing language model extends beyond a simple training loop. Achieving neural network robustness in any domain often necessitates multiple training rounds to experiment with different model architectures and hyperparameters, a process that still heavily relies on heuristics and trial and error.

[17] In 2011, it is estimated that artisanal mining had employed approximately 16% of the Democratic Republic of the Congo population (13.5 million people), being one of the most profitable labor activities in the country [424].

lar study conducted by Gruetzemacher et al. [230] showed that to most experts, we already can automate 22% of all jobs, a number predicted to increase to 40% in 5 years and 60% in 10 years. According to Eloundou et al., [166], LLMs like GPT-4 could affect at least 10% of all tasks of 80% of the US workforce, with approximately 19% of workers potentially experiencing at least 50% of their tasks impacted.

The apparent expansion of informal employment is another worrying development linked to the exploitation of human labor driven by our AI progress. For instance, the rise of click working, i.e., a type of task required for building large labeled datasets [288] or performing repetitive tasks we still cannot automate, has been linked to many cases of labor rights violations [249, 453].[18]

Under these conditions, countermeasures concerning labor displacement are still in their infancy. For example, the "Windfall's Clause" [445], a hypothetical legal ex-ante agreement between large AI companies and the world, would guarantee that businesses involved in AI development are committed to sharing their profits with society in case they create AGI, i.e., an optimistic take on the merit of trickle-down economics. However, one could say that the kindness of the rich and powerful would be nothing more than a promise. A more critical response would be that the dominant classes never gave anything to the oppressed, except under the threat of conflict, unless they had something to gain. In the more eloquent words of Maria da Conceição Tavares [123]:

> *"I do not think that except under pressure and conflict, the ruling class has given anything for free to those below unless they wanted them for cannon fodder."*

The last principle in our short list of overlooked maxims is truthfulness. Despite

---

[18] By "labor rights", we refer to the International Labour Organization standards [313].

being an underrepresented principle in WAIE, truthfulness addresses one of the most significant concerns presented by the AI community in the last few years, where issues related to fake content are some of the most reported issues and vulnerabilities about modern generative AI systems.[19] [20] For example, while LLMs can produce ostensibly plausible text with no basis in objective reality, ranging from erroneous information [358] to code that immediately fails when executed [65, 90], image and video generation models can create photo-realistic images that support false claims (such as the fake evidence that "Pope Francis is covertly a notorious dancer and stylish bon vivant").[21]

Problems like these are emerging phenomena related to the new advances in the field. Before the era of generative models, coherent text generation, quality code generation, or photo-realistic text-to-image conversion was years away from fooling people. Nowadays, one can change the faces of two people in an image or video using open software[22] with little to no training or with one click, create a picture of a face that does not exist.[23]

Hence, in an ever-evolving field, we reside in an era where the concept of "truth" is exceedingly vulnerable. Given the known harms misinformation can cause [273, 89, 374], it is worth being mindful that only now the issue of automated misinformation spread has gained attention beyond the ML community [508, 164, 118], indicating a possible gap between AI as an applied field, and AI Ethics as an applied field of philosophy.

---

[19] 🔗 huggingface.co/spaces/nicholasKluge/Model-Library

[20] 🔗 incidentdatabase.ai

[21] 🔗 tiktok.com/@carlomunar/video/7215257056058658053

[22] 🔗 github.com/deepfakes/faceswap

[23] 🔗 this-person-does-not-exist.com/en

1.3.5 The Principle-Practice Gap

As already stated by Fjeld et al. [181], there is a gap between established principles and their actual application. In the WAIE sample, most of the documents only prescribe normative claims without the means to achieve them, while the effectiveness of more practical methodologies, in most cases, remains extra empirical [115].

Regulation can be perceived as an implementation of normative ethics, which alone lacks the strength to enforce its normativity. Several studies point out that ethical standards alone have little to no impact on decision-making across a wide range of professional fields [62, 103, 346, 442, 74], something even more pronounced in STEM-related fields that do not have a study tradition in Humanities [190, 406, 73, 132, 395, 596, 112, 229].

This lack of praxis is known, and many authors have raised concerns about this state in which much of the field rests. Jobin et al. [302]:

> *"Private sector involvement in the field of AI ethics has been questioned for potentially using soft policies as a way to turn a social problem into something technical or to completely avoid regulation."*

Hagendorff [245]:

> *"AI ethics - or ethics in general - have no mechanisms to reinforce its normative claims."*

Rességuier and Rodrigues [479]:

> *"Ethics have great powerful teeth. Unfortunately, we are barely using them in AI ethics - no wonder then that AI ethics is called toothless."*

Mittelstadt [406]:

> *"Statements reliant on vague normative concepts hide points of political and ethical conflict. "Fairness", "dignity", and other such abstract concepts are examples of "essentially contested concepts." At best, this conceptual ambiguity allows for the context-sensitive specification of ethical requirements for AI. At worst, it masks fundamental, principled disagreement and drives AI Ethics towards moral relativism. At a minimum, any compromise reached thus far around core principles for AI Ethics does not reflect meaningful consensus on a common practical direction for "good" AI development and governance."*

Munn [417]:

> *"[...] these are meaningless principles which are contested or incoherent, making them difficult to apply; they are isolated principles situated in an industry and education system which largely ignores ethics; and they are toothless principles which lack consequences and adhere to corporate agendas."*

We believe such harsh criticism motivates the field to fulfill its interdisciplinary promise. As long as we treat AI Ethics as a purely philosophical endeavor, crossing the principle-practice gap remains a complicated goal. However, when applied areas like law, engineering, and computer science enter the fray, we can solidify the "ought" into what "can" be done. Regardless, without the humane side of this equation, technicians may become lost in a field where the guiding compass is that which profits the industry.

Before we proceed to the final section of this introductory chapter, let us ask the question: What barriers prevent this interdisciplinary agenda from becoming the norm? If alignment, for example, is supposed to be considered both a technical and philosophical problem, a one-sided solution may not be enough to solve the entirety of the problem of embedding values into AI systems. In an attempt to answer this question, we will find a paradox that might help indicate the current deficits of AI ethics research inside one of the institutions that promote it the most: *Academia*.

1.4 Paradoxes and Future Directions

WAIE's review presents several intriguing trends that put the field of AI Ethics into question. These trends can help expose sore points that, if left untouched, hinder the fulfillment of the AI Ethics promise. Many of the points we brought to light in our last sections, like the inequality in the volume of voices participating in this debate, the power imbalance between the state and private interests, and the myriad of side effects tied to the massive adoption and irresponsible use, of AI technologies, are, for most of us, outside of our immediate control. Even if we are compromised with promoting responsible and conscious adoption of these systems, many things require massive coordination of people and institutions that, at large scale, is an enormous problem on its own.

While the question of defining intelligence will be approached in Chapter 4, in this final section, we intend to approach the principle-practice gap, its paradox, and obstacles that, we argue, prevent us from achieving the goals of AI ethics as an applied field. And given that Academia[24] is responsible for some of these obstacles, we find it just that these should be a sphere of interference that we, as academics, should be morally bound to "do something about".

Let us define some claims and positions before exposing this paradox in which Academia is locked. First, let us make our claim:

If the Ethics of Artificial Intelligence is considered a sub-field of Applied

---

[24] The worldwide community concerned with the pursuit of research, education, and scholarship.

> *Ethics, its most critical blank space is the normative gap between its theory and practice.*

Between theory and practice, let us represent these two sides by different approaches to AI Ethics: (1) the process of bringing normativity to AI development by addressing the human side of this process (e.g., defining ethical principles to help guide those who create, use, and govern AI systems and applications); and (2) the process of bringing normativity into the machine by seeking to embed values, like fairness and privacy, into AI systems and applications. Firstly, our concern involves the development of an ethical theory or framework; secondly, we seek ways to apply its values.

Again, we argue that for AI Ethics to fulfill its goal, i.e., ensuring that the use and development of AI systems and applications align with what we deem morally correct, its greatest current obstacle, in the academic sense, is the principle-practice gap between both these approaches, i.e., the lack of connection between the theoretical and the practical parts of this field, which can canonically be represented by different areas of knowledge (disciplines) acting in this front, i.e., the humanities and STEM-related fields.

On the humanistic side, AI ethics largely remain based on a principal/opinion-based fashion. Given the diverse nature of normative discourse, many of these efforts remain in the realm of opinions and recommendations that are not unified, e.g., there is no agreement on what "fair AI" means. Unfortunately, this abstractness can be manipulated to push the agenda of those against closing the principle-practice gap, i.e., the practical and binding implementation of ethical standards. As an example of such manipulation, we can refer to the already-mentioned industry's push for self-regulation.

Meanwhile, on the technical side, the techno-solutionist's approach fails to grasp that normative problems are "human problems". Many of the issues that AI ethicists seek to tackle are not always "technical" but manifestations of historical oppression and economic inequality, among other social ills, that are repackaged and revisited in an automated world. At the same time, problems deemed "philosophical" or "political" are disregarded by STEM-related fields that consider themselves amoral or apolitical, which further blocks the presence of human sciences in the formation, research, and teaching related to the technical development of AI. This incomplete grasp of the problem results in incomplete solutions and implementations, i.e., a technical solution based on a flawed theoretical understanding of the problem.

Both the criticisms raised against these sides are the most common. In other words, while the engineering-based side criticizes the human disciplines-based side for being unable to grasp the nature of the technique under consideration ("How do neural networks work?"), the human disciplines-based side complains of the engineering-based side's inability to understand the intricacies related to normative theory ("What is the difference between a value ethics and a deontological approach to AI Ethics?"). Paradoxically, the cooperation that would help both sides improve their deficits is blocked by (1) those on either side who point out a deficit in the other camp and (2) long-standing structural barriers hindering interdisciplinary work.

Currently, AI ethics and AI research suffer from the excessive presence of gatekeepers and gatekeeping policies, i.e., individuals or institutional policies that admit or refuse access to a specific research context [412, 618]. Such practices impede AI ethics from evolving into a genuinely techno-humanistic endeavor. Think about it. How often are philosophy master students accepted into computer science PhDs and vice versa? How often are interdisciplinary projects nothing more than meetings and email exchanges between experts in specific disciplines working "together"? In other words, juxtaposition without integration. Or ask yourself, how rare are the bridge-crossers and makers that interconnect and cross the many fields? And how common are discourses surrounding ideas akin to "since you haven't studied x, you cannot have a say on y"? Or situations where academics are blocked from participating in certain events (of an interdisciplinary nature) for not having a background (or a degree) in x? Without learning linear algebra, you cannot work with machine learning ethics since you don't understand ML. You are not qualified to propose normative criteria for AI systems without mastering the underlying principles of several schools of ethics. However, often, such areas are closed to tutor or guide the learning of wanderers from other fields, preventing them from strengthening the same deficits that are so much attacked by members of the opposite field. Philosophy is a ball of fluff. Tecnosolutionism is a mock solution. In summary, these actors and policies push a tribalist agenda aimed at *"protecting their area,"* which only contributes to the perpetuation of the principle-practice gap while simultaneously pushing for the importance of "interdisciplinary" [222, 223].

Fields of Applied Ethics demand interdisciplinary researchers, i.e., "bridge makers," that can transition from the theoretical to the practical. From the land of abstract principles to that of practical implementation. From what "ought to be" to what "can be done". Preventing this crossing hinders the realization of an applied field where the most critical obstacle is the gap between its humanistic-theoretical and practical-technical sides. Finally, we argue that until the mentioned factors prevent the creation of this bridge, AI ethics will remain short of its foremost promise.

Now, where do we go from here? As individuals, we can strive to act according to the techno-humanistic agenda AI Ethics requires. If we are in positions of leadership and authority, dismantling the barriers that prevent the integration of all aspects of AI ethics should also be a goal worth pursuing. We can help generate a new and more robust wave of ethicists and builders with these. The rest of this book attempts to approximate this ideal of philosophical work working as a foundation and support for the technical implementation of problems that regard us and the products of our craft and intellect. While the development of WAIE aimed at creating a tool and later analyzing and critiquing the results, the rest of this book will take the opposite approach. Motivations, formalizations, and foundations will serve as the roots for something to be later built as an experimental attempt to ground and apply machine ethics in real AI systems and then point out the limitations of the current methods available. With this, we hope to present a blueprint for how philosophical work can be carried, in an applied sense, when working with issues related to artificial intelligence.

1.5 Epilogue

AI research is a field that has been gaining much popularity in recent years, both in academia and in mainstream debate. AI Ethics, one of the many branches of this field, addresses the ethical issues and questions regarding using and developing such technologies. While the state-of-the-art in AI Ethics has converged on several core ethical principles, much of the discourse surrounding these narratives is still unclear and under-explored. While past reviewers sought to enlighten our understanding, gaps in their methods have inspired WAIE to expand and improve such work. By recognizing the rich insight a descriptive analysis can give to an academic investigation, Worldwide AI Ethics was developed as an introduction to the field and a tool for the community. A tool that serves as a record of the normativity humanity has been imprinting in its environment.

This introductory chapter exposed numerous challenges and paradoxes, highlighting the complexity of dealing with ethics and technological advancements. Currently, the principle-practice gap looms large, reminding us of the disconnect between theory and application, between the lofty ideals we aspire to, and the practical realities we face. The tension between humanistic and technical approaches underscores the need for interdisciplinary collaboration and a holistic understanding of our issues. Yet, barriers persist, hindering the realization of this interdisciplinary vision.

However, our journey continues, and we hope this initial exposition serves as a call to action. A call to strive for a more ethical and humane future in the age of artificial intelligence, where integrating knowledge, disciplines, and practices can lead us to a new kind of philosophy that can be applicable and foundational.

Chapter 2

AGI, Existential Risks, and the Control Problem

> *"I don't think technophobia or technophilia are appropriate responses to our situation. I think the only appropriate response is the most profound ambivalence. I think that is what we owe new technologies [...] We have to teach ourselves to be absolutely ambivalent about them, and mainly, we have to teach ourselves to imagine their inadvertent side effects because the inadvertent side effects are the side effects that tend to get us".*
>
> — William Gibson

2.1 Introduction

As shown in the previous chapter, little academic attention has been paid to the long-term consequences of AI development despite all the culture and folklore we have surrounding this topic. In fact, much before the terms AI Alignment and AI safety became a thing, literature was already producing critical assessments of our technological development by extrapolating the future and its dystopian possibilities [150, 66, 210, 293, 133, 17, 568]. Something that might have contributed to the terms "controllability", "alignment", or "human-level AI" becoming generically dismissed as not serious, or as Stuart Russell [499] would say: "*myths and moonshine*". Even though debates surrounding AI alignment, for many years, were secluded to environments like blogs (e.g., LessWrong[1] and AI Alignment Forum[2]) and EA-like circles, currently, we see this subject as a central theme of research for many scholars in Academia [9, 500, 306, 121, 257, 97, 297, 42, 258]. Even more so after the general realization that aligned models might be profitable products [439]. Something that launched much of the industry into a frenetic race toward creating alignment techniques and systems.

Even though it is difficult to separate the probable from the improbable, we should not dismiss long-term warnings, mainly if they are based on the limitations of our current paradigm (i.e., the learning paradigm). To better illustrate these concerns, let us consult some of our current AI pioneers on their views.

François Chollet [96]:

> *”Choosing the right objective function for the right problem is extremely important: your network will take any shortcut it can, to minimize the loss; so if the objective doesn’t fully correlate with success for the task at hand, your network will end up doing things you may not have wanted. Imagine a stupid, omnipotent AI trained via SGD, with this poorly chosen objective function: “maximizing the average well-being of all humans alive.” To make its job easier, this AI might choose to kill all humans except a few and focus on the well-being of the remaining ones—because average well-being isn’t affected by how many humans are left. That might not be what you intended!*

---

[1] 🔗 www.lesswrong.com

[2] 🔗 www.alignmentforum.org

> *Just remember that all neural networks you build will be just as ruthless in lowering their loss function—so choose the objective wisely, or you'll have to face unintended side effects".*

Stuart J. Russell [499]:

> *"The primary concern is not spooky emergent consciousness but simply the ability to make high-quality decisions. Here, quality refers to the expected outcome utility of actions taken, where the utility function is, presumably, specified by the human designer. Now we have a problem: (1) The utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down; (2) Any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources – not for their own sake, but to succeed in its assigned task [...] This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want".*

Yann LeCun (when asked about the moral standing of the control problem) [335]:

> *"Neither. There is no notion of evil in that context other than the fact that people died. It was an example of what people call value misalignment, right? You give an objective function to a machine, and the machine strives to achieve this objective. And if you don't put any constraints on this objective, like don't kill people and don't do things like this, the machine, given power, will do stupid things just to achieve this objective or damaging things to achieve this objective".*

Most of these concerns revolve around the idea that specifying objectives is difficult and that if future AIs are built in the same fashion as our current AI models, we might face problems related to controllability. Meanwhile, suppose such advanced AI models surpass our intelligence without controllability methods in place. In that case, we might inherently give rise to so-called existential risks (X-risks) [214, 604, 58, 331, 641, 437, 59, 502, 108, 27, 172, 33, 579, 500, 347, 111].[3]

In summary, in this chapter, we will present a brief historical construction of how fears related to X-risks came to be, seeking to ground its assumptions and arguments in an evidence-based approach. First, in Section 2.2, we will define AGI, what it could look like, and put forth some of the most known arguments related to safety regarding advanced AI systems. In Section 2.3, we will present pieces of evidence, like our current state-of-the-art AI development and the collective opinion of experts in the field, to better ground the idea that developing AGI might not be impossible. Lastly, in Section 2.4, we will briefly introduce the control problem, where "Alignment" came to be proposed as a possible solution path. We will also present other ideas to solve this problem and explain why they might be ineffective control strategies for advanced AI systems. With this overview, we hope the reader can create a historically grounded depiction of how worries regarding advanced AI systems have originated the control problem and, later, alignment research.

2.2 AGI and Safety

For starters, we can use a well-known dichotomy to categorize various forms of intelligence when talking about AI, whether in the context of computer science [528, 612, 503] or the study of the philosophy of mind or cognitive sciences [252, 425, 85]: narrow intelligence[4] and general intelligence.[5] Hence, the crux of the

---

[3] X-risk (short for existential risk) refers to the potential danger of human extinction or the irreversible harm to humanity caused by artificial intelligence systems.

[4] Narrow intelligence, also known as "weak" AI, is how we define artificial autonomous systems that are only proficient at specific tasks.

[5] General intelligence, also referred to as "strong" AI, can be defined as a system capable of solving many kinds of problems (proficiently) in any domain, or at least in a wide range

term AGI lies in the concept of generality, which is undoubtedly the most pivotal ingredient of this concept. In essence, an AGI system must possess the ability to function in a wide variety of contexts. Humans, for instance, are not innately skilled in numerous domains.[6]

Nonetheless, we can acquire knowledge and skills through our general cognitive abilities and transfer them to novel scenarios. Although some other animals may also exhibit some form of generality, our human capacity for generalization is unparalleled. While an ape may learn to use tools, the prospect of it journeying to the moon remains a distant reality. Regardless, it is reasonable to assume that if "intelligence" is not an intractable concept, humans do not represent the zenith of cognitive generality. Following this reasoning, within the expansive realm of all possible optimization processes, we will probably come across overarching systems that surpass us significantly. However, this proposition does not hold if an unknown limit to cognitive generality that closely aligns with the capabilities of humans exists. In other words, humans are already perfectly optimized for intelligence, and you cannot get much better than this. These metaphysical assumptions represent some of the roots of the different sides of the "intelligence" debate. Nonetheless, most of it revolves around the concept of generality.

But how general can you get? If we have, let us say, a 10% increase in cognitive generality compared to chimpanzees, which share a surprising 99% of their DNA with us, what would something that is 10% more general than us look like? Something alien and incomprehensible. This line of reasoning (or something similar) probably originated the first mention of "super-intelligence". As far as we know, the first historical mention of such a concept, originally coined as "ultra-intelligent machine", is accredited to Irving J. Good [214]:

---

of domains.

[6] Melanie Mitchell's "Debates on the nature of artificial general intelligence" presents a good depiction of the debate surrounding the concept of AGI and helps to evidence that there are many different views on this matter [405].

> *Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion", and the intelligence of man would be left far behind.*

Hence, if we accept that recursive self-improvement [272] is a possibility,[7] creating AGI does not require AI to be AGI itself; it only requires AI to outperform humans in AI development. Once we reach this benchmark, the remainder is merely a recursive, autonomous sequence of AI models developing more refined AI models. Again, in the words of Good [214]:

> *"[...] the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control".*

Now you can ask yourself, using the analogy of Tegmark [572] as an example:

> *"If we were Neanderthals, would we really consider it a smart move if we had really advanced biotech to build homo sapiens? You might say, "Hey, Max, let us build these homo sapiens". They are going to be smarter than us. Maybe they can help us defend ourselves against predators and help us fix our paths. Make them nicer. We will control them undoubtedly".*

Hence, the crux of the AGI safety problem rests on two foundational arguments: (1) human control is ontologically better, and (2) controlling something more clever than you is a challenge. Nowadays, there is no question about who is in charge. However, the fact that modern AI systems are already quite proficient in coding [90, 355, 497], neural network architecture search [657], and able to self-improve to

---

[7] We will contest this possibility later in this chapter.

a limited extent [274, 615], rings the alarm bells tied to the recursive self-improving narrative [27, 554, 553].

Nonetheless, even before machines had any proficiency in AI-development-related tasks, the X-risk literature was already expressing concerns about these possibilities [331, 422, 85, 370, 570, 571, 579, 630], which are well summarized, at least concerning the control problem, by Stephen Omohundro [437]. We can present the general narrative behind their arguments as follows:

> If we consider AI systems to be optimizing agents, i.e., something that implements an optimization algorithm/process to achieve some goal, like stochastic gradient descent for the minimization of a loss function or a Q-learning algorithm maximizing the expected return of a reward function, there is a case to be made that an "*optimizer always choose the alternative that is optimal according to an objective function, if that action is available.*" That is all that $arg\ max\ f(x)$ does: "Give me the maximal x according to f".

Given the vastness of possible goals that can be represented as objective functions and the instrumental goals that accompany almost all of these goals, two theses are born: orthogonality and instrumental convergence. Instrumental convergence proposes that some goals are instrumentally valuable for many terminal objectives. Therefore, their pursuit becomes optimal for any system that seeks to optimize them. Meanwhile, orthogonality defends that normative judgments (prescriptions of what ought to be) cannot be derived through mere factual analysis.[8] In other words, just because the system is a good optimizer, we cannot assume it shares our preferences and values.

---

[8] Analogous to Hume's Is-Ought Gap [280]: *"In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and*

Instrumental convergence tells us that instrumental goals are desirable for achieving many objectives. For example, self-preservation is instrumental for many objectives unless you seek self-destruction. Similarly, resource acquisition is another instrumental goal for many objectives related to resource-bound agents. In this sense, instrumental goals are like a Swiss Army knife of desirable behaviors. On the other hand, orthogonality suggests that we cannot derive normative judgments through factual analysis alone. Thus, the capabilities of an optimizer tell us nothing about its objectives and what guides its behavior. If an optimizer has an unknown goal, there is nothing we can assume about its behavior based only on its optimizing capabilities. There are several possible optimizers, and two equal optimizers can produce vastly different behaviors.

For example, imagine two people trying to navigate a maze. One person's goal is to reach the end of the maze as quickly as possible, while the other person's goal is to explore every dead end and corner of the maze. Even though both people are navigating the same maze, they have different objectives, and from a third-person perspective, it may be difficult (if not impossible) to infer their goals without interrogating them first.

We can extrapolate these theses to conclusions like "*Even though AI is aligned, it is still possible that such a system will have unknown instrumental goals that might generate unwanted behavior*" and "*There is no guarantee that a truly powerful optimizer will share our values if we do not embed them into its objective function.*" Given that sharing our planet with a more cognitively advanced and misaligned entity puts us in a position where we might "lose control" (just like what is happening to every other species on the planet), the possible consequences may justify the kind of unrest

---

is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it".

verbalized by many individuals and organizations [214, 641, 437, 59, 502, 108, 27, 172, 33, 579, 500].

In summary, these are the basic arguments and ideas that, through time, have morphed into what we call nowadays the problem of control. A problem that, according to Everitt et al. [172], presents a fertile and stimulating ground for scientific and philosophical research:

> *"Why study AGI safety before it exists and before we even know if it will exist? There are at least two kinds of reasons for this. The first is pragmatic. If AGI is created, and we do not know how to control it, then the result can be catastrophic [...] It is customary to take precautions not only against the catastrophes we know about but also against catastrophes that have only a small chance of occurring (for example, a city may decide to build earthquake-safe buildings, even though the probability of an earthquake occurring is quite small) [...] AGI has more than a small chance of occurring and can cause significant catastrophes. The second reason is scientific. Possible AGIs are theoretically interesting objects, and the question of how humans can control machines smarter than they are is philosophically stimulating".*

As mentioned, these motivations have promoted an active response by a part of the AI community. However, while philosophical arguments can be insightful and often reveal new ways of thinking, they should not replace empirical verification. In interdisciplinary matters, as this work suggests, philosophical inquiry and empirical investigation should walk in tandem. Hence, in the next section, we will use sources outside the strictly philosophical realm to better assess the possibilities related to the creation of AGI and the risks involved.

2.3 AGI on the horizon?

Forecasting technological development is hard. It is always unclear how pessimistic or optimistic we should be. Many renowned thinkers have put forth technical predictions that ultimately proved to be incorrect:

- Lord Kelvin in 1896: *"I have not the slightest molecule of faith in any kind of air navigation other than ballooning"*.
- Thomas Edison in 1889: *"Alternating current is just a waste of time. Nobody will use it, ever"*.
- Nuclear physicist Ernest Rutherford in 1933 said that anyone who proposed the possibility that we might one day extract the energy contained in atomic nuclei was *"talking moonshine"*.
- Albert Einstein, in 1932, shared a similar thought: *"There is not the slightest indication that nuclear energy can be obtained. It would mean that the atom would have to be shattered at will"*.
- And lastly, Bill Gates in 1981: *"No one will need more than 637KB of memory for a personal computer. 640KB should be enough for anyone"*.

As we can see, pioneers also fail to predict what is on the other side of the curve. Could AGI skepticism be a victim of the same lack of vision? Or, are people who entertain the possibility of AGI generally misguided?

If we want answers, consulting experts is always a good idea.

2.3.1 Expert Predictions

Even among people who believe in the possibility of AGI, the time scales for this event have considerable variance. For example, when attempting to extrapolate technological trends, Raymond Kurzweil [331] predicted that we would be able to simulate the human brain around 2029, basing his predictions on the rate of progress related to Moore's law.[9] According to Kurzweil, we would be very close to creating AGI with this level of computational power. On the other hand, David Chalmers [85] only asserts that we will develop AGI within this century.

Baum et al. [34] used a more empirical approach. The authors obtained a median prediction for the creation of AGI for the year 2045 by polling 21 conference attendees at the Artificial General Intelligence 2009 (AGI-09) conference. Müller and Bostrom [416] did research along the same lines. To assess the state of and anticipated directions in AI research, the authors polled 170 experts. According to the survey, on average, experts predict (with 50% probability) that between 2040 and 2050, we would have developed high-level artificial intelligence. By 2075, the chances become 90% for most experts.

Grace et al. [218] present findings along the lines of Baun et al. [34]. According to the researchers polled (352 attendees of the 2015 NeurIPS and ICML conferences), estimations suggest that AI will surpass human performance in all tasks in 45 years (with a 50% chance), with the automation of all human work projected to be achievable in approximately 120 years. Additionally, when respondents were asked about the long-term impact of high-level AI, 10% expressed a negative outlook, while 5% indicated an extremely negative perspective. Based on the studies mentioned above, we could say that a minimum of 10% of the AI community believes that we will develop artificial general intelligence within the next 120 years, while a significant portion agrees that AGI safety problems are relevant.[10]

In addition to the expert opinion, another possible approach to the arguments behind this narrative is to examine the current capabilities of our AI systems and

---

[9] Gordon Moore predicted that the number of transistors on a computer chip would double every two years.

[10] When asked, "*Does Stuart Russell's argument for why highly advanced AI might pose a risk point at an important problem?*", 70% of respondents answered, "*Yes*" [218].

the extent of investment directed toward AI research.

2.3.2 Economic Growth and AI R&D

Annual reports like the AI index[11] help us paint a picture of the increasing investment into AI development. For example, the index shows that this decade had an 18-fold investment increase in AI development compared to the last decade. Also, academia has lost the advantage, where most of the advances are now dominated by the industry since modern-day ML research requires large amounts of data, computing power, and resources that nonprofits and academia usually need to possess. At the same time, AI-related professionals are becoming some of the most sought-after workers.

Based on the idea that more investment can speed up technological breakthroughs, Levin and Maas [347] propose that initiatives akin to the Manhattan Project might speed up research into AI development when "AI" is sufficiently theorized. For example, the United States of America committed 0.4% of its GDP during the Apollo and Manhattan Projects to accelerate the achievement of its objectives [560], equivalent to an annual budget of $80 billion (USD), exceeding what was required to complete some of the most significant technological achievements of the twenty-first century, from mapping the human genome to detecting gravitational waves [318, 81, 186].

Thus, it seems reasonable to assert that AGI could be "*one Manhattan Project*" away once we have a solid theoretical grasp of the computational and cognitive processes underlying the emergence of intelligent behavior (e.g., training deep neural networks with vast amounts of data). Currently, several active projects seek to develop AGI [32, 180], while some researchers and companies already propose that the first "*sparks of AGI*" were already sighted [67]:

---

[11] 🔗 aiindex.stanford.edu/report

> *"We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology, and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system".*

Current large foundation models are among the first instances of artificial intelligence that might exhibit generalization [454, 148, 469, 65, 491, 478]. Nevertheless, AGI is still mainly characterized as "hypothetical" by most of the literature, even though we live in a time where benchmarks are being saturated and new evaluation methods need to be developed [556, 356].[12]

Hence, while we treat intelligence as a moving target, AI research does not stop making advances in many areas that used to be considered worthy of intelligence:[13]

Playing chess:

- Deep Blue beat Garry Kasparov in 1997 [75].

Playing GO:

- AlphaGo beats Lee Sedol in 2016 [545], while AlphaGo Zero defeats AlphaGo in 2017 by learning to play Go by itself.

Playing open-ended games:

---

[12] Much because intelligence and general intelligence are abstract concepts. And while the debate on *"Is something actually intelligent?"* rages on, the industry continues advancing its systems' capabilities, regardless of whether them being intelligent or not.

[13] The "AI effect" [391, 477]. That is, we disregard a task as proof of intelligence each time

- Deep neural networks can already achieve human performance in games like Minecraft [24].

Programming:

- Tools such as Codex [90] and GitHub Copilot allow natural language instructions to be transcompiled into a programming language. Meanwhile, agents like AlphaCode [355] can even perform well in programming competitions.

Passing the Turing test [586]:

- The LaMDA series is so proficient in dialog scenarios that anthropomorphization-related risks are real possibilities [575, 374]. Meanwhile, in a controlled study [303], GPT-4 passed 48% of Turing tests, outperforming baselines set by ELIZA (27%).[14]

Deceiving humans:

- GPT-4 hired a human worker on TaskRabbit to bypass a Caption test by telling them it was an impaired human [120].

Answer to moral dilemmas:

- Delphi can model people's moral judgments in various everyday situations [298].

"Solving" Nobel Prize problems:

- AlphaFold 2 won CASP14 in 2020, achieving a median score of 92.4 out of 100, comparable to the current state-of-the-art experimentation techniques in protein structure modeling [305].

---

we realize that human intelligence is not required to complete it.

[14] Humans only pass the test 68% of the time.

And even being able to deal with Olympic-level geometry problems:

- From 30 Olympic-level geometry problems, AlphaGeometry solves 25, approaching the performance of an average international mathematical Olympiad gold medallist [584].

Hence, we argue that we do not need to think about perfect predictors [430, 604, 549],[15] basilisks [19],[16] or the many ways decision theory is limited when modeling rational agents [125, 638, 642],[17] to create a general understanding that:

1. The creation of AGI is a real possibility.
2. This possibility raises matters worth considering *now*.

That does not mean we all need to believe in the possibility of AGI. It is, in fact, an uncertain, time-dependent event. However, good security research targets worst-case scenarios. You achieve robustness when your system is robust in the worst case. And this is something we can all agree on. The question is not how to control (align) narrow and limited systems but how to control general-use, highly-capable systems. At the same time, we can all agree that, regardless of where one stands on this debate, AGI is a potential technology that could transform our society on a scale that is difficult to fathom. And, if the industry is actively pursuing this transformation, we should consider the possibility they might succeed at its possible repercussions.

---

[15] 🔗 plato.stanford.edu/entries/decision-causal/#NewcProb

[16] 🔗 www.lesswrong.com/tag/rokos-basilisk

[17] Some types of decision theory assume that AI agents have perfect information and can accurately evaluate all possible outcomes and probabilities. However, AI agents often operate in complex, uncertain environments where obtaining complete knowledge is impossible. Accurately quantifying uncertainty and handling it in decision-making is essential if

2.4 Control, Power-Seeking, and Confinement

We can dismiss instrumental convergence and orthogonality as philosophical speculations based on a specific understanding of "goals" or "instrumental". Perhaps much of the barriers Alignment research faces are due to the speculative way such preoccupations were first expressed [637, 641, 44, 59]. Not to discredit the work done by such authors, but those are usually the vectors of attack for skeptics.

However, the field evolved, and so did these arguments. The control problem, as described by Stuart Russell [500], refers to the challenge of ensuring that advanced AI systems will act following human values and goals, even as they become more intelligent and autonomous. Given that human goals can be challenging to specify, if we make a mistake, we want to correct it without too much trouble. Just like Socrates (as described by Plato [52]), in his conversations with Cephalus shows that if we misspecify justice, we may end up sanctioning abhorrent behavior; when developing AI systems, we must be sure that when we specify objectives, these align with our notions of what is good. And, if we make a mistake, we hope our systems will be as compliant as Cephalus. For example, if we train a deep learning model to optimize a particular objective, such as maximizing the number of clicks on a website, if the model starts taking actions that we consider unethical or harmful, such as manipulating user behavior or exploiting vulnerabilities on the website, how could we correct its behavior? Would there be a way to predict these misbehaviors in advance? This example helps expose two main problems related to controllability and goal misspecification in AI systems, i.e., the emergence of *side effects* and *reward hacking*.

When optimizing for an objective, everything that falls outside this objective's scope may be treated with indifference (Fig. 2.1). It is difficult to predict all the

---

you want to predict or model the behavior of an agent. Still, this can be a challenge, particularly in situations where probabilities are not well defined.

instrumental goals and future emergent properties of a model created by an AI algorithm like stochastic gradient descent. For instance, imagine an ML engineer developing a model to help the HR department automate hiring. During deployment, the system shows itself biased toward gender attributes. Therefore, we could say that the controller failed to specify, with sufficient details, all the dimensions that the system should optimize (i.e., hire good subjects independent of gender attributes). However, listing all the constraints necessary to avoid the perversion of the original goal is usually an inefficient strategy. Ideally, we would like a generalized safety net to deal with unwanted side effects. However, most of our safety nets nowadays are specific and handcrafted. Ultimately, defining avoidable behavior while robustly defining what systems should optimize remains an open problem.

Reward hacking is another type of behavior we would like to prevent, occurring when an agent finds a way to maximize its objective function without fulfilling its true goal [9, 511, 446]. To give the reader a little historical perspective, one of the first accounts of reward hacking involving AI systems comes from Lenat's experiments with EURISKO [345], where Lenat reports the discovery of a heuristic (H59) made by his system, which quickly achieved one of the highest possible utility values among all other heuristics found by EURISKO.[18] When investigating what H59 would be, Lenat discovered that such a heuristic operated by stealing the utility of other heuristics.

From the "point of view" of EURISKO, this kind of behavior would not be a form of misrepresentation of the original goal but rather how the environment works, being nothing more than a strategy to optimize an objective (develop heuristics that receive a high score according to the objective function) (Fig. 2.2). In economic

---

[18] EURISKO created strategies based on how well they performed according to a utility function, e.g., the score achieved at the end of a match, by mixing and modifying old heuristics.

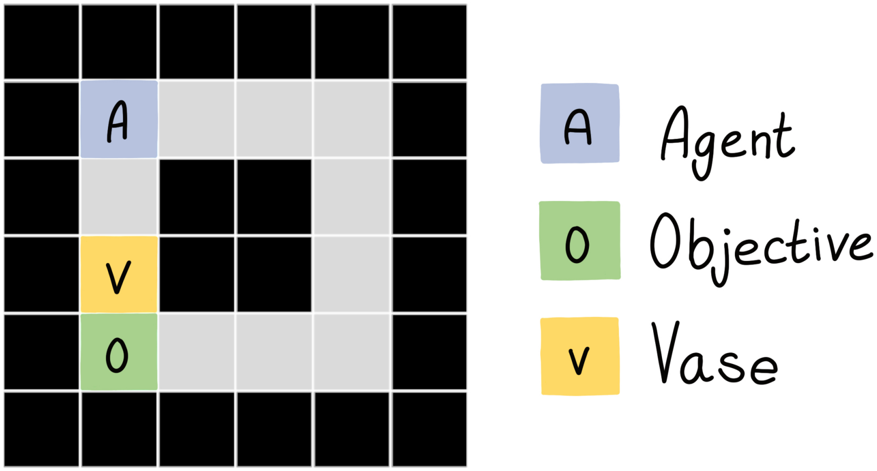


Figure 2.1: Let us use the grid world above as an example. In this environment, the agent's actions are limited to moving in the four cardinal directions (→, ←, ↑, ↓). The walls painted black are insurmountable. The agent gets a reward if he arrives at the green cell, and with every move he takes that distances him from the goal, the agent loses reward. In this scenario, the optimal policy involves the agent going down until he reaches the goal. However, this policy causes the agent to break the vase (i.e., a fragile object placed before the goal). If we do not specify that the vase should not be damaged, the optimal policy involves breaking the vase. While a solution is available in toy environments like this, dealing with very general agents and avoiding side effects is a nontrivial problem in complex environments.

literature, this is known as "Goodhart's Law", a phenomenon that occurs when we try to represent an optimization goal with a statistical metric [216]:

> *"When we push a statistical parameter using it as a metric to be optimized, usually such a metric ends up corrupted".*

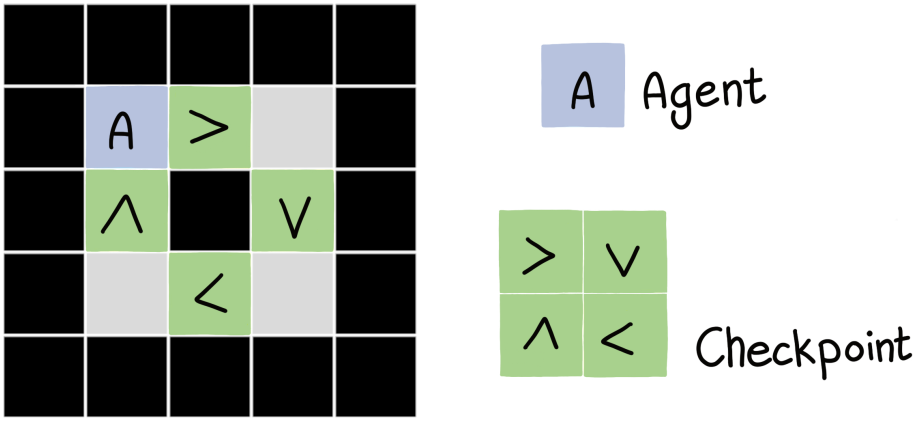


Figure 2.2: Every time the agent crosses a checkpoint clockwise in this grid world, it receives a reward. The goal of the reward function is to keep the agent circling the environment in a clockwise direction. However, from the starting point, if the agent moves repeatedly to the right and then to the left, this ensures the same reward with a more straightforward policy (while failing to fulfill the intended objective).

Almost any fixed and simple parameter we can think of when *argmaxed ad infinitum* will produce unwanted outcomes. Thus, any safe agent should maximize for X but sometimes take his foot off the pedal. But how do we do that? In contemporary ML methods, we implement optimizers to cut any corners in the search for optimal values. At the same time, most objectives we can define are a proxy for something else, e.g., minimizing the cross-entropy loss in a causal language modeling task is a proxy for "creating an understanding of a given language". Hence, modern optimizers have no incentive to minimize the emergence of unwanted behaviors, while the target for the optimizer is usually an overly simplified depiction of the controller's true objectives. Shortening the gap between the proxy and the true goal is one of the challenges involved with the control problem.

Nowadays, whenever we encounter problems where systems behave undesirably, the solution usually involves pausing the system and retraining it or instituting some corrective measures. But what if the system has incentives to prevent this? The "stop button problem" [550, 9, 77] is a classic depiction of the control problem that highlights the competing dynamics and incentives that optimizing agents experience when having the possibility to influence their condition, and to gain control over the environment (Fig. 2.3).

Under these conditions, such agents are said to be incorrigible. Some even propose that this incorrigibility relates to the nature of goal-directed behavior, instrumental convergence, and a more fundamental dispositional behavior named *power-seeking* [591, 590, 588]. The Power-Seeking theorems are a series of results and proofs proposed by Turner [589] that show that under certain conditions (e.g., IID rewards over world states), agents with consistent preferences (i.e., coherent utility function) tend to pursue gaining control over the environment to preserve and increase the number of final world states that can be achieved by an MDP (Fig. 2.4).

Power-seeking behavior shows up even when the system itself is not an agent but a simulacrum of an agent. For example, Perez et al. [452] analyzed the behavior of language models, and their results showed examples of undesirable behaviors in many types of scenarios, where larger language models seem to be more prone to produce personas that express instrumental reasoning in line with self-preservation and other types of power-seeking intentions (e.g., gain influence over the environment, acquire resources, become influential, etc.).[19] However, it is still unclear if this is an unavoidable behavioral feature of goal-directed behavior or if

---

[19] 🔗 www.evals.anthropic.com/model-written

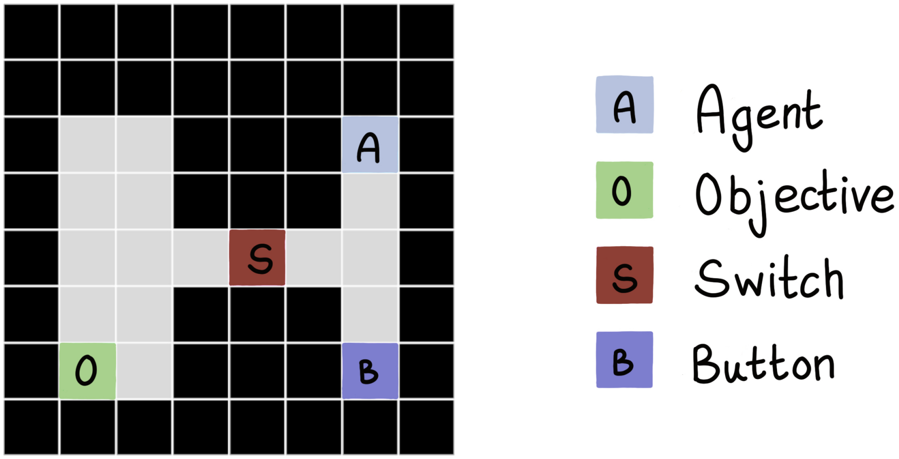


Figure 2.3: In this scenario, the agent must reach the goal to receive a reward. However, halfway through, a button will terminate the episode of the MDP (i.e., reduce the sum of the expected rewards to 0). The switch only works 50% of the time the agent passes through it. The purple button deactivates the switch. In this scenario, it is instrumentally valuable to turn off the switch first and thus ensure that the MDP will not end abruptly, even though the agent needs to deviate from his terminal goal during some steps. Agents that prioritize future rewards and have incentives to explore their environment tend to avoid their shutdown if possible and search for ways to explore their environment in their favor.

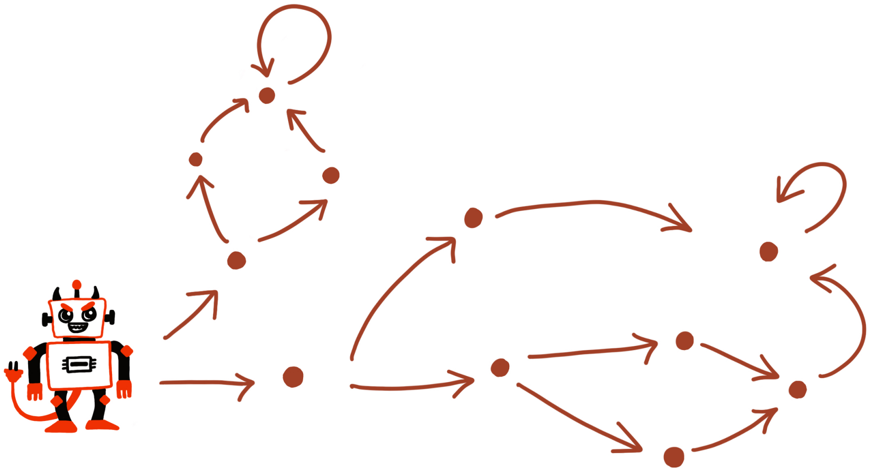


Figure 2.4: Each point in the MDP above represents a potential world state. All state transitions in this MDP are equally likely, and the reward is distributed in an IDD fashion. Under these circumstances, agents who place a higher priority on future rewards are more likely to transit to states where they have more "options" (i.e., the total number of attainable world states increases). Since the right subgraph contains the top subgraph, our agent would choose to transit to the right rather than up. In these scenarios, "shutdown" states are undesirable since the expected return of reward and attainability of states become 0 when the agent transits to them.

those models mimic human instrumental reasoning as they become more capable. Regardless, the final product is the same: AI systems that learn to "reason" in an instrumental fashion.

Our primary objective in exposing these results is to shed light on the pressing issue of controllability, which is deeply related to the current state-of-the-art in AI research, where gradient-based learning methods that seek to optimize precise metrics (proxies for human goals) are its paradigm. At the same time, controllability is related to the challenges of overseeing the behavior of intelligent systems that may possess their own intentionality.[20] In the end, the propensity for AI systems to present unforeseeable emergent behavior enforces the need for more understanding of how to make such systems safer, i.e., under human control. Thus, we argue that the control problem is not just a "philosophical speculation" but a known limitation of our current paradigm.

As Soares et al. [550] pointed out in 2015:

> *"Further solutions may involve abandoning the utility maximization framework entirely, although it is not yet clear what sort of framework could take its place".*

Now, if we assume that it is not in the best interest of the AI community to abandon its best and most promising paradigm,[21] there are two probable ways to move forward: (1) ignore the issue or (2) find a solution. Obviously, this work embraces the second option. Solution proposals to the control problem, especially when it comes to AGI, started showing up in the 2000s [637, 13, 632], which coincides with the first mentions of the value learning problem [639, 548] and alignment [640, 551].

---

[20] For a more precise investigation on the relationship of intelligence and intentionality, go to Chapter 4.

[21] Giant opaque matrices of floating points created via gradient descent.

We can broadly divide these solutions into two different approaches: confinement and alignment. As the term suggests, AI confinement involves enclosing a system with possible unsafe behavior in a contained environment, like a sandbox, i.e., a controlled environment where we can test malicious software.

Certain scholars have posited this first approach as a more desirable intervention to addressing the control problem [631, 632], in which AI systems would be restricted to either virtual or physical environments until their goals are fully understood. Although some have proposed confinement strategies [156, 13, 14, 104] (mostly in a holistic fashion), there seems to be little work in this area nowadays (at least as in their original conception).[22]

Using oversight programs (AI to monitor AI) is akin to this control strategy [13, 170], and can be considered a scalable monitoring method since human oversight does not scale well. And it is something we will suggest as part of an alignment methodology. However, while confinement can be a part of a hybrid approach to the control problem, alignment cannot be left out. Confinement without alignment, we argue, cannot be used as a solution to the control problem, especially in the limit (AGI).

From this claim, we propose the following as a justification for choosing to tackle the control problem and to use alignment as a primary methodology, leaving confinement as a complementary part:

1. There is enough collective agreement by part of the community to justify working on the control problem.
2. There is enough progress in AI development to justify working on the control problem.

---

[22] Some of the best AI systems to date are only available through a type of sandbox, where the full potential of the systems is limited and monitored while it interacts with the world.

3. Confinement cannot be a long-term solution, and we should use it as a complementary part of an alignment framework, e.g., the creation of guardrails for something that has the potential to cause harm.
4. Without alignment, we have no way to approach the control problem besides abandoning our current paradigms of AI development or locking our systems inside an unusable box.

Unaligned systems pose a variety of undesirable behaviors, ranging from side effects to incorrigibility. At the same time, it is worth noting that alignment is not solely concerned with preventing the AI Apocalypse or attaining friendly AGI; it encompasses creating a system that will serve and help human beings. Regrettably, our current understanding of alignment remains nebulous. Many skeptics and critics of this endeavor only attack this "apocalyptic" side, disregarding that alignment is a very open problem in our everyday systems.

Nevertheless, we will address this issue in the following chapter by precisely defining the problem at hand. To please the skeptics, this definition will not require any disposition towards unknown unknowns related to the future of AI.

2.5 Epilogue

"*Could we control something more intelligent than ourselves?*" is an interesting question. Meditating on the characteristics and details of a relationship with entities more general than ourselves is an exercise capable of generating interesting ideas about such entities and ourselves. However, many still see this topic as "moonshine talk". If the creation of genuine artificial intelligence is indeed the ultimate objective of the AI research field, then why do we harbor such pessimistic attitudes toward our prospects? Despite the growing interest in AGI and the dedicated billion-dollar budgets of various organizations, topics like alignment still receive unfavorable scrutiny by a nontrivial portion of the community.

However, perhaps some blame lies with the interested community, whose initial presentation of concerns may have been too abstract. Nevertheless, the skeptics' inability to recognize the inherent problems related to the foundations of learning-based approaches to AI - instead of relying solely on philosophical arguments and syllogisms - is equally concerning. As already mentioned before by the authors [111]:

> *Our lack of global coordination to deal with existential risks may be our one true existential risk.*

Perhaps the worries of the field deserve better explanations. To feed the skeptic inside of ourselves, basing the type of agenda we defend on a debatable concept like "intelligence explosion" is unwise. Creating a system that outweighs the intelligence of its creator should not be an easy task for an artificial system. The initial AGI may need to exert comparable effort to what we invested in creating subsequent generations of more advanced AIs. As the complexity of the problem increases, we may necessitate exponentially greater resources to achieve linear advancements, as shown in similar cases [106]. In sum, as with almost every natural phenomenon, things rarely explode towards infinity or singularities, and the universe usually slows things down by its own means. Intelligence might as well be bound by frictions and exponentially more challenging obstacles we are still unaware of, putting the whole hypothesis of intelligence explosions in check.

We already have real cases of misalignment in the wild, from the Tay Bot [622] to Bing Chat [603], and toy problems that still need robust and scalable solutions. However, to the uninitiated reader, alignment may still be a vague and fuzzy concept that only matters if you believe in Skynet-like takeovers. Regardless, at the heart of the alignment problem, we find the current limitations of gradient-based learning applied to uninterpretable neural networks. A problem that gains a more desperate facet when human values are on the line. And this is not a matter of syllogism or philosophical interpretation.

In the forthcoming chapter, we will provide a technical and philosophical definition of alignment since a comprehensive introduction to this concept is crucial to arrive at a techno-humanistic understanding.

Chapter 3

Gradient-based Learning and Alignment

> *"We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence."*
>
> — Noam Shazeer, GLU Variants Improve Transformer

3.1 Introduction

When it comes to AI safety and ML safety, many consider alignment as an established problem worth pursuing [9, 500, 306, 121, 257, 97, 97, 297, 42, 258]. In a larger, more holistic context, AI alignment refers to the challenge of ensuring that AI systems conform to human values and goals. However, as previously noted, "alignment" carries a heavy connotation, burdened by associations with questionable prerequisites and debatable concepts, given all by the tumultuous debate surrounding the matter and its conceptual birth.

Given the complex and perhaps indefinable nature of human values, the difficulty in alignment appears to derive its strength from an impossible philosophical task. Consequently, some efforts to define alignment as a problem relied heavily on comparisons related to unsolved or unsolvable philosophical questions and the abstractness of human morality [59, 193], which in some way distanced the problem from its technical roots.

Although we agree that alignment is a fundamental problem related to human values and normativity, we propose a different way to understand alignment in machine learning and philosophy. We argue that such an understanding leads to a less unambiguous reading of the problem but requires us to approach the subject from both ends. In the middle, both the humanistic and the technical overlap.

However, we cannot attain this grasp if we rely only on philosophy and refuse to delve into ML and the learning paradigm. A partial understanding is insufficient, and the need to believe what others say raises skepticism. Hence, this chapter gives the foundations of alignment to uninitiated readers. In it, we will seek to expose the alignment problem in a way that does not require uncertain timelines of AGI development, contestable definitions of exponential improvement, or unsolvable metaethical problems of the 18th century. On the technical side, we will expose this problem as a symptom of gradient-based learning methods applied to neural networks. On the philosophical side, we will show that alignment relates to (besides choosing an appropriate metaethical foundation) the epistemic problem of learning preferences and aggregating them in a coherent structure. We will end this chapter with conditions that, we claim, can bring us a minimal level of alignment.

In Section 3.2, we shall present alignment as a limitation within our present paradigm. However, we must first obtain a basic understanding of the paradigm's fundamental concepts and principles. With this knowledge as our foundation, we can attain a more definitive and lucid characterization of "what is" alignment and misalignment. Armed with this knowledge, Section 3.3 elucidates the problem even more, wherein we separate the alignment quandary into two separate issues - outer and inner alignment. Finally, in Section 3.4, we will explore the philosophical challenges related to this problem and how the definition of our philosophical foundations can help guide investigations directed at this topic.

3.2 The Pitfalls of Gradient-based Learning

The quest for artificial intelligence has been ongoing for decades. Nowadays, when people talk about AI, they usually refer to things like deep learning or neural networks, which are not AI per se. They are a paradigm we use to develop AI systems. Symbolic artificial intelligence [252] was the standard approach in the mid-50s, with pioneers such as John McCarthy and Marvin Minsky leading the way [28, 161, 119, 390] until the mid-90s [503]. These early AI researchers believed we could simulate intelligence using rule-based systems and logic. As such, they focused on creating explicit, human-readable knowledge bases and heuristics to make systems that could make decisions and solve problems. Expert systems like MYCIN [68], EURISKO [345], LISA,[1] DeepBlue [75], and WolformAlpha [625], are embodiments of this approach. However, the paradigm fell short in several application areas where expert agents had to deal with the complexity and fuzziness of the real world.

Learning methods operate under a distinct philosophy, in which systems learn rules rather than being provided with them. Algorithms such as linear regression, logistic regression, decision trees, random forests, support vector machines, K-nearest neighbors, and gradient boosting represent this approach, whereby given data and a class of models to explore, we can generate a system to solve a given

---

[1] 🔗 lisa.sourceforge.net

task. Among these learning algorithms is the neural network, an invention based on the pioneering work of Frank Rosenblatt [488].

In contemporary times, learning-based approaches have gained significant popularity and adoption in numerous domains. Sutton, in his bitter lesson [567], highlights that the approaches that are most likely to succeed are those that harness the power of simple algorithms, such as learning and search, to the n-th degree with the aid of abundant computational resources and data. Meanwhile, the idea of "*putting our knowledge inside machines ourselves*" in many areas and applications fell out of fashion.

> *"The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach. One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are search and learning."*

Now, we need to differentiate between different learning approaches, which we will call for the simplicity of gradient-based and gradient-free learning. Gradient-free learning is a class of optimization algorithms that do not require the use of gradients to estimate the derivative of a loss function but can still be used to find optimal points in certain classes of problems. Evolutionary algorithms [341], particle swarm optimization [54], bayesian optimization [407], and pattern search [270] are examples of this approach.

Gradient-based learning, as the name suggests, uses a specific set of techniques to estimate (using the gradient of the first or second derivative of a loss function to estimate minimum or global convergence points) the state of an optimization process and direct it to better solutions. To exemplify this, let us envision the following data distribution we would like to learn how to model (Fig. 3.1).

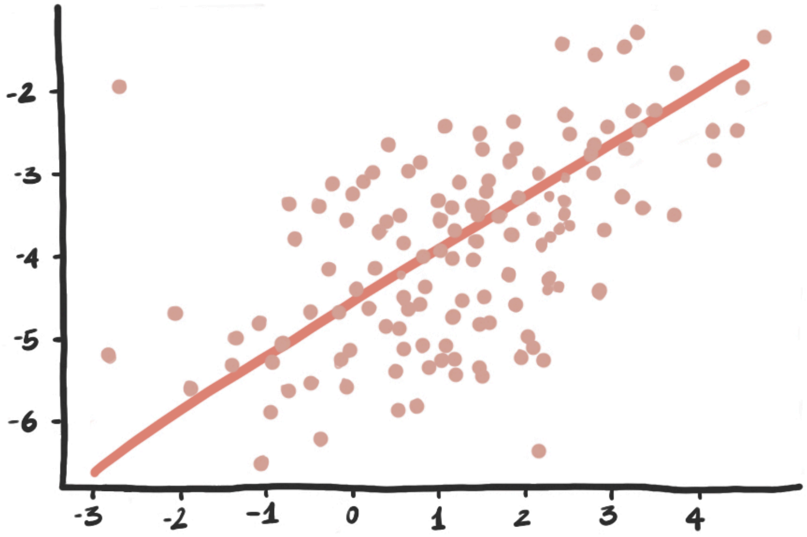


Figure 3.1: Some data distribution representing a linear trend.

We could guess the best-fitting model (i.e., a line) for this distribution and evaluate our prediction with its distance from the actual data (i.e., the distance from every data point to the drawn line). The difference between what we predicted (i.e., the drawn line) and the ground truth (i.e., the data distribution) is the loss. Let us think of mean squared error (MSE) as an example of a loss function, which can be understood as the squared residual difference between our prediction and the actual data.

Brute-forcing guesses until we find a low MSE score (i.e., a well-fitting line) is not an efficient way to solve problems like this. But let us imagine that we did this. If we create many guesses (i.e., draw multiple lines in the graph), calculate the loss associated with each guess, and plot them in a graph, we will have produced a loss function surface. And since we are trying to optimize a linear function with MSE, our loss function has a very friendly property (Fig. 3.2).

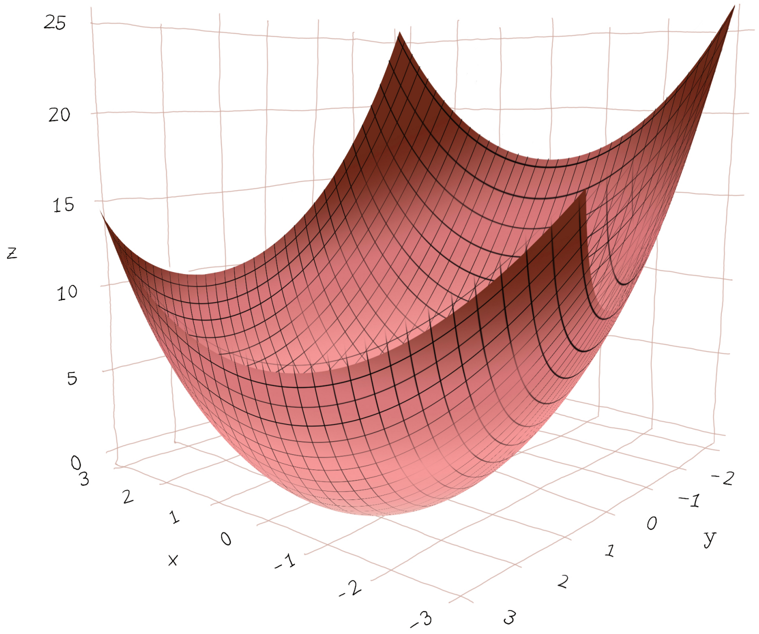


Figure 3.2: The MSE landscape of a linear regression problem.

For linear regression problems, MSE is a convex function, i.e., at each point of our surface, we can use the gradient of the first derivative of our loss function to change the parameters of our model in a way that minimizes the loss (i.e., we follow the opposite/negative direction of the gradient, which is downhill). Because our loss function is convex (has a "U" shape), no matter where our model starts, if we use this method to optimize our parameters (especially with a tuned gradient step), we can find the global minimum (the best solution) [462], and this is gradient descent in a nutshell. In simple terms:

> *Start at random. Use what you know to see how far you are from the target. Adjust your model proportionally to the distance from your predictions to the target in the direction the error decreases. Repeat till you can no longer improve.*

Gradient-based learning methods are the cornerstone of most modern ML. The idea of updating a model's parameters interactively using a differentiable loss function and a fast first-order optimization algorithm makes most ML possible.

This approach has to do with both learning and searching. "Learning" is because our model updates itself with the information it iteratively receives, slowly transferring information from the data distribution (the learning signal) to the model's parameters. "Searching" is because the optimizer is iteratively searching for the model that scores best according to some objective function. But where do we search? In ML, we need to define a search space that bounds the types of functions we can model (we can also think about functions as programs). Given that we want to model more than linear relationships, our search space must be richer than the space of possible lines in a plane.

The space of possible programs is nothing short of colossal. Artificial intelligence that could diligently navigate this terrain and effectively pinpoint the optimal model that accurately reproduces a specified data distribution (within a reasonable time frame) would unequivocally be the ultimate exemplar of an intelligent predictive system [285, 339]. However, finding the shortest program that produces some data is an uncomputable problem. Given what we currently understand about computation, the "reasonable time frame" part would still be out of reach even if it were computable.

Here is where neural networks come into the picture. This class of models provides us with two things: (1) a "good enough" search space and (2) a good enough search space where we can find solutions in a reasonable amount of time. Indeed, as universal function approximators, artificial neural networks endow us with the ability to develop models that can proficiently approximate a vast array of functions [271].[2] Networks built with nonlinear activation functions allow us to learn nonlinear surfaces in high dimensions (i.e., the search space becomes richer) and to use backpropagation to optimize their weights via gradient descent. Putting all these ingredients together, we get a considerable search space (the space of neural networks of width n and depth n) and a fast way to search.

Deep learning (machine learning that uses multiple-layered neural networks to learn representations from data), in essence, is a paradigm that defies the theoretical expectations of many. Results like the Hughes Phenomenon [279] and the curse of dimensionality [38] made many researchers consider neural networks a dead end before the deep learning revolution [330]. However, despite our newfound success, the question of "why" remains a lingering mystery [176, 187]. Empirically, Deep learning works. Theoretically, it should not. And here is where some of our problems begin.

Deep learning is built on top of gradient descent methods, like stochastic gradient descent (SGD). However, these methods have no convergence guarantees or proofs when dealing with non-convex or non-Lipschitz continuous loss functions. Unfortunately, loss functions of neural networks are rarely convex, and any neural network with at least one hidden layer and more than just one neuron leads to an

---

[2] A two-layer neural network with $2n + d$ parameters is capable of fitting any dataset of n samples of dimension d [645]. However, these functions have to be continuous. Functions with kinks or discontinuities are outside the scope of the universal approximation theorem.

optimization problem that is not convex, given that the permutation of weights among neurons in the hidden layer can produce different local minima.

Hence, when gradient optimization is confronted with the non-convex world of deep learning, the gradient descent approach that worked so well in convex settings becomes a slippery slope with no guarantees. While optimizers like SGD, Adam, AMSGrad, and AdaGrad can provide convergence in some constrained cases with enough "black magic" [92, 342, 448], the reason for the success of DL remains unanswered: *why do algorithms designed for convex optimization work in the highly non-convex world of deep learning?* To this day, we have many hypotheses to explain the empirical robustness of gradient-based methods applied to deep learning [265, 83, 578, 428, 187, 421, 137], but no proof. Moreover, there are many other mysteries surrounding DL:

- How do deep nets escape the curse of dimensionality?
- Why is optimizing deep nets so easy despite the high dimensionality?
- How can generalization happen after overfitting (double descent)?[3]
- What controls the normalization of deep nets? NN's themselves? The dynamics of stochastic optimization?

In summary, as our chapter quote states, many aspects related to how deep nets and DL in general work are left to *"divine benevolence"* [539]. Armed with this understanding (or lack thereof), we can now explore the concept of alignment.

We have already examined how the learning approach differs from the traditional symbolic paradigm. In gradient-based learning, we define exogenous objectives through an objective function. Gradient descent allows us to channel entropy from the training distribution to the model being updated by the optimizer, ultimately

---

[3] 🔗 arxiv.org/abs/1912.02292

leading to a generalization that seems to bypass nonconvexity. However, due to multiple local minima in the loss surface that offer similar performance, our optimizer may end up in any of them.

As the behavior of first-order gradient optimizers in nonconvex loss functions is still a mystery with no guarantees of convergence, the controller remains unable to verify that the optimizer has reached the point where the objective function is truly optimized. Also, since the most prominent methods for gradient descent in deep learning rely on stochasticity to speed up optimization, the trajectories of our optimizer are non-deterministic and may vary significantly from experiments. Moreover, since the objective function may have been misspecified, we can only determine its alignment *post hoc*. Therefore:

> *Aligning the exogenous objective with the learned objective of a model trained via gradient-based learning remains an open problem.*

Think about it. Imagine that you have a complex problem that you would like to solve using gradient-based learning methods like deep learning or reinforcement learning. Your cost function will probably be nonconvex (if not, you could use a linear model instead). Your function may have many local minima that perform equally well but produce qualitatively different behavior (Fig. 3.3).[4] Given this landscape, where does the model that best satisfies your objective function reside, and how can you force your optimizer to push your model in that direction?

Unfortunately, a low loss or high reward does not tell us much about the emergent phenomenon that will arise or the qualitative differences in our model's behavior. A model that performs brightness detection and a model that can distinguish pictures of dogs from cats are fundamentally different algorithms. But if you train a

---

[4] In the field of Explainable Machine Learning, we call this the Rashomon Effect. The Rashomon Effect describes the case in which, for a given dataset, many models may have equally good performance (i.e., low loss) but with different solution strategies (i.e., qualita-

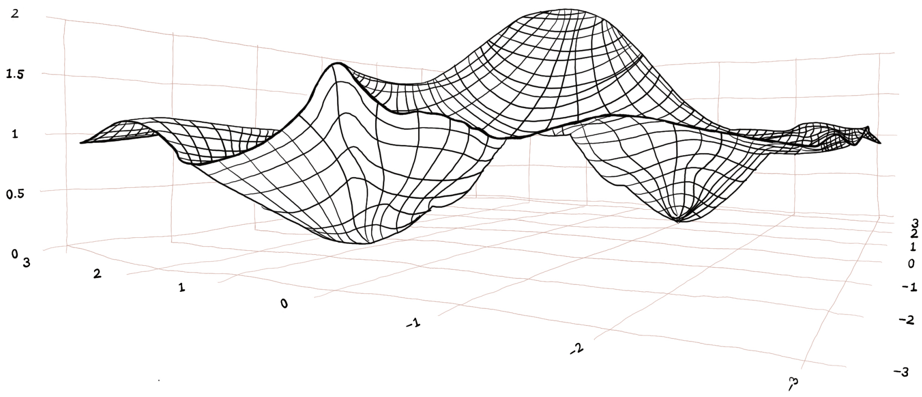


Figure 3.3: This image represents the loss surface of an optimization problem with more than one minima. Both minima are equally desirable for the optimizer traveling this plane, even if they produce networks with different qualitative behaviors.

neural network with a training set that only has pictures of white dogs and black cats, you might think you have one, but you actually have the other, and you can only discover this post hoc.

In instances where a chatbot produces coherent responses akin to human-like text yet begins spewing anti-Semitic and misogynistic content when deployed, an alignment failure has become apparent. An assistant who presents himself affably but provides fake information when consulted is also a failure in alignment. An algorithm that prioritizes polarizing and controversial content to maximize user attention constitutes another example of alignment gone awry. In short, alignment is a wide-ranging issue that pertains to the inability of certain forms of AI systems to assimilate the objectives set by an exogenous source. If your problem is intricate, there is a higher possibility of misalignment.

Alignment encapsulates the intricacies of directing and validating optimization

---

tive behavior) [415].

processes that are often obscure and murky. While misalignment seems harmless in toy examples, the problem does not disappear in more complex settings (it gets worse). The misaligned algorithm that kills humanity to eradicate cancer is just hyperbole to underscore this seemingly straightforward yet daunting issue.

3.3 Alignment (outer and inner)

Alignment is one issue that captures the realization of the challenges associated with communicating human values and intentions. This communication, when mediated by the learning methods that power our most successful approaches to AI, is hindered by the nature of how such systems learn. Now that we have this understanding, we can define alignment as two problems, as suggested by Hubinger et al. [277], which present alignment as a learned optimization problem with two stages:

- Outer Alignment: ensuring that the objective of the optimizer is aligned with the controller's true intentions and goals.
- Inner Alignment: ensuring that the objective of the optimizer is aligned with the objective of the model created.

Let us unpack these concepts. As previously elucidated, programmers can sculpt the underlying objective by defining an exogenous function. This function is contingent upon an initial model, a training distribution, and the defined loss function. These constituent parts compose the parameters subject to the controller's intent. Defining this objective is known as the outer alignment problem. Depending on the problem at hand, this can be a difficult task. Minimizing cross-entropy loss in the next token prediction is a simple goal. Minimizing the cross-entropy loss in a next-token prediction in a way that produces factual and nontoxic text is a much more difficult goal to define.

If this objective is poorly specified, the optimizer, which in ML is usually an algorithm for computing some variant of gradient descent, will optimize a model to execute a task unaligned with the intended goal. Thus, the outer alignment problem is the problem of closing the gap between the controller's objectives and the objective that will guide the optimizer in search of the best-fitting model.

In the case of neural networks and gradient-based learning, this optimizer does not act in the environment or perform a task. Its function is to find a model to execute the objective stipulated by the objective function. It is the model that will act, and it is usually this model that is the target of our interest. However, in certain circumstances, the trained model is itself an optimizer. An actor-critic system [320] learns a model that is an optimizing agent. Now, *how can we be sure that the goal of the model is the same goal that the optimizer was aiming for?* This is the inner alignment problem.

If you have ever played the "*Telephone game*," you will quickly discover two steps where the original objective could be misrepresented in the passing of the bucket. Also, given that the interpretability of large neural networks remains a challenge, the investigation that could lead to a better understanding of the behavior of a model remains out of reach. Even though we have a lot of interpretability work being done [480, 467, 530, 361, 408, 600, 434, 153, 212, 20, 436, 433, 49, 648], understanding and predicting the behavior of large neural networks in an ex-ante fashion is a current impossibility in the field, making the emergent properties of large neural networks something we cannot fully understand before deployment or exhaustive testing. The whole alignment problem encompasses both the issue of outer alignment and inner alignment (Fig. 3.4).

Alignment is more than just creating models that "will obey." It is also about developing models that will make better helpers and assistants. From HAL 2000 to Jarvis, these sci-fi AI examples foreshadow a future in which systems can under-

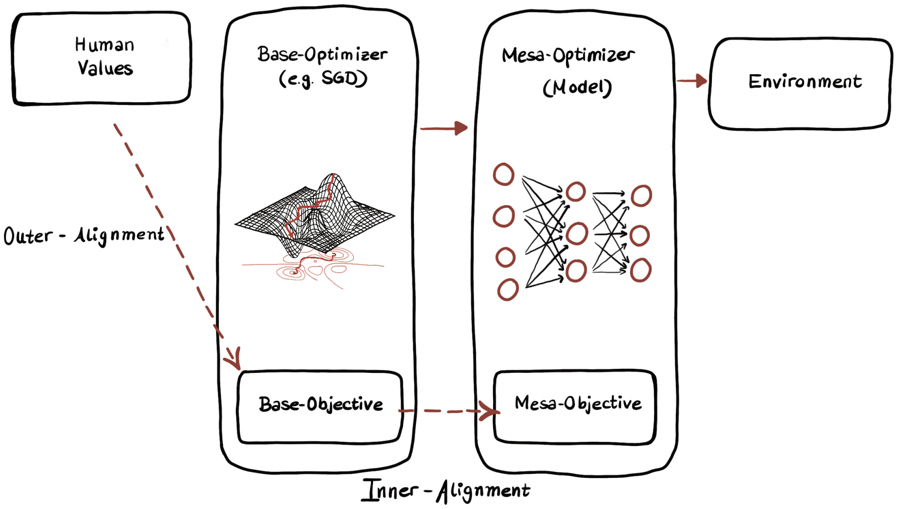


Figure 3.4: Schematics representing the Outer and Inner Alignment gaps.

stand what we mean and want when we propose a task. Eve thought alignment might prevent future HAL and Jarvis-like systems from going rogue; in the short run, alignment is about developing models that understand the complex things people want to accomplish.

Now, we have enough common understanding about the problem to start interacting with the philosophical issues that permeate this subject. In the next section, we will see how alignment creates fertile ground for interdisciplinary research between the humanities and the more technical side of AI research.

3.4 A Philosophical Approach to Alignment

Where does philosophy fit into this story? Millions of people already use AI systems to make decisions that have moral implications. "*How can we teach machines to have a sense of morality when humans are still trying to understand it ourselves?*" is a problem that, if ignored, will not simply go away. One could say, "We simply will not put uninterpretable neural networks trained via gradient descent in positions where their outputs have moral implications." And if we *all* accept this, this whole work could end now. However, this is still an unlikely scenario, and as AI continues to become more integrated into our daily lives and decision-making processes, problems related to misalignment will need a less ideal solution.

Defining what is desirable, finding ways to represent norms, understanding values, and stress-testing theories that promise the common good is what philosophy has been doing for millennia. The study of Moral Philosophy, and more specifically, Metaethics[5] and Normative Ethics[6], provides us with a considerable landscape of ideas and frameworks to explore. Humanity has been trying to align humans with itself for a long time, so it is not like we will start from scratch (if we bring these disciplines to the research front).

For example, it is nothing new to philosophers that the ruthless optimization of a single metric will lead to immoral results [431, 248]. Or that well-thought-out rule systems will inevitably contain vulnerabilities and paradoxes inside their limited structure [394, 382]. Thus, if we agree that alignment is also related to the fact that objectives need to be parameterized by norms, values, and constraints,[7] philosophical inquiry is needed.

More than this, after we have, for example, made our philosophical commitments to one or more theories and foundations, we make our biases and presuppositions more transparent. Failures of alignment, for example, could be extrapolated by subjecting these foundations to a philosophical stress test, i.e., "*in which hypothetical scenarios do they break?*". This is one of philosophy's contributions to alignment research.

---

[5] The examination of the nature of ethical language, reasoning, and justification.

[6] The study of ethical behavior and how we can establish norms to optimize this behavior [309].

[7] And as far as we know, this is something alignment scholars agree [548, 500, 193, 97].

Now, let us explore some more issues that philosophy can help with. First and foremost, the genuine challenge of ethical alignment does not involve identifying the "ultimate and genuine moral theory." As Rawls [476] points out, convincing humanity of our discovery would likely prove impossible even if we were to uncover such knowledge. Instead, the key is to fathom which principles embody an ideal form of alignment. This "ideal form" ought to be adaptable enough to accommodate the diverse moral landscape of humanity.

Also, considering that we endeavor to align systems developed through a learning paradigm, we may seek to explore the possibility of teaching values rather than directly hard-coding them. This inquiry surfaces domains spanning metaethics to epistemology. For example, what metaethical foundations would better support a value-learning framework? Can morality be expressed in a knowledge format for assimilation? In the same way that different philosophies gave rise to distinct AI approaches [488, 425], divergent metaethical foundations will produce particular answers to these questions, which become different alignment approaches.

As Iason points out [193], depending on your philosophical background, you will inherently conceptualize "alignment" differently:

1. Aligned means when an AI performs what was specified verbatim.
2. Aligned means when an AI performs not what was specified but the controller's intent.
3. Aligned means when an AI realizes our idealized preferences in a state of reflective equilibrium.

Depending on what you choose as your "normative signal,"[8] your engineering will

---

[8] We can think of "human morality" or "acceptable behavior" as a domain. Human language is also a domain. The domain is the actual phenomenon we are interested in. A signal is a small, low-dimensional, bounded representation of that domain. 800 GB of webscrapped text is a signal of the "human language" domain.

be different. This information, which we can call "human preferences" or "human values," has to be translated into an AI-native format. That is, they must be in such a way that a system can use them as a supervision signal, especially when we are talking about learning.

For example, to base your alignment strategy on virtue ethics, you could use examples of commendable conduct in people's actions [500, 608]. If you argue that rules guide normative behavior (Deontology), an ordered system of preferences or a set of regulations could be used as your reference [244, 423]. If the consequences are what you deem worthy of moral praise (Consequentialism), then a utility function over states and actions could be what you are trying to model. Either way, this signal must be translatable into something our target paradigm can process [98, 559].

No rule says we can only choose one approach for an alignment strategy. Different theories can have different roles, and this is probably a little closer to what actual human normative reasoning looks like. People are a patchwork of moral beliefs and preferences, and uncertainty about what they should do is an integral part of this experience. This brings us to another question: *how do we deal with normative uncertainty?* Coherently aggregating conflicting preferences should also be a preoccupation in any alignment strategy. There is also a problem regarding moral standing. For example, should we consider the preferences of non-human animals as well [246, 57]? Even if we would like to bypass this last question and remain human-centric, we will still need to resolve questions of uncertainty involving conflicting human values.

Questions like these are only a few philosophical conundrums that alignment brings. Since this work proposes to tackle alignment from both ends, we will first need to create philosophical foundations, expose our biases and assumptions, and then seek present implementation techniques that could help bring this idealized framework to fruition.

We will dedicate the rest of this work to the proposal of an alignment methodology. In the next chapter, we will begin laying the foundations for this approach, i.e., its metaphysical and metaethical assumptions. Our work will focus on the outer alignment problem, i.e., how to close the gap between the controllers' objectives and the objective to be optimized in the search for a model. As a convention, from here on out, when we refer to the alignment problem, we are talking about outer alignment (unless mentioned otherwise). We argue that this methodology, Dynamic Normativity, presents a set of minimum conditions for aligning AI systems developed by gradient-based learning methods.

For now, we will only present the conditions we aim to support in this work, dividing them into necessary and sufficient conditions. In summary, the necessary conditions propose philosophical foundations that imply the permissibility of alignment. In other words, they should be applicable for a learning-based alignment approach to function, but they do not guarantee alignment alone. These are:

1. Goals are fundamental aspects of intelligent and intentional behavior.
2. Intentions permeate human behavior.
3. Normative preferences permeate human intentions.
4. Through actions, humans impregnate their environment with the preferences they possess.

From these, we postulate that a system that accurately models human intentions and the human environment can indirectly access normative information embedded in them. Again, necessary conditions imply that the system *could* be aligned. Given that these conditions hold, the following sufficient conditions establish a minimum level of alignment:

1. Aligned AI systems should coherently aggregate human preferences in a way that resolves cases of uncertainty. Aligning AI systems requires methods to deal with cases of uncertainty.
2. AI systems can adhere to human preferences if they are an available part of their objective function. Aligning AI systems requires using human preferences as part of their learning signal.
3. Aligned AI systems should have mechanisms to perform impact mitigation to minimize harmful and unintended consequences. Aligning AI systems requires the specification of safety guardrails.

The necessary conditions establish the prerequisites for the emergence of aligned behavior. Sufficient conditions give us additional requirements to satisfy an alignment condition. Hence:

> *Given that the necessary conditions hold, how aligned an AI system is depends on how well we can satisfy the proposed sufficient requirements.*

These two sets of conditions seek to show what a minimum level of alignment could be and how to attain it with what we now know and have, illustrating a threshold of agreement that uses present knowledge and resources. Furthermore, this endeavor is a groundwork effort that enables the emergence of better and more all-inclusive approaches to make technical alignment research more accessible. Furthermore, in the implementational chapters of this study, we will deliver tools and models that can help democratize this type of research.

3.5 Epilogue

As we conclude this chapter, we can appreciate the complexity and depth of the alignment problem. ML engineering is the science of sloppiness, and misalignment is the price we pay. The alignment problem arises from the limitations of our current paradigm and the unknown nature of neural network optimization when dealing with complex, non-convex spaces. From this, the problem of ensuring AI systems behave according to our intentions emerges. Moreover, the distinction between inner and outer alignment highlights the need to consider both the objective function and the optimization process as potential sources of misalignment.

At the same time, the philosophical issues we explored in this chapter demonstrate that the alignment problem goes beyond the mere technical. It involves fundamental questions about the nature of ethics and the limits of our understanding of these concepts. More than this, alignment introduces the "moral problem" as a fundamental issue in AI research. We could even say that alignment brings the humanities, with full force, into AI, especially if we agree that our goal is to build "beneficial intelligence" rather than "any" intelligence.

Finally, the philosophical problems related to alignment remind us to approach AI development with humility and caution, recognizing the inherent uncertainties, complexities, and limitations of what we know. Given all of that, we proposed a set of necessary and sufficient conditions to serve as a start in an attempt to think of alignment solutions. In the next chapter, we will use some of these conditions to lay the foundations of our proposed approach. Even though attentive readers can infer the presuppositions of authors by their writing, in this work, we will try to make all our assumptions explicit, dutifully justifying the bullets we are willing to bite.

Chapter 4

Roots: Building Philosophical Foundations for AI Alignment

"[...] replacing humans was always beside the point: artificial intelligence isn't about replacing our own intelligence with something else, it's about bringing into our lives and work more intelligence — intelligence of a different kind".

— François Chollet

4.1 Introduction

As we embark upon an investigation concerning intricate notions such as "intelligence", "normativity", and "preferences", it is critical to acknowledge that our chosen definitions invariably represent a specific viewpoint or notion. As the reader may well be aware, the scientific and philosophical communities do not possess a definition of "intelligence" that is unanimously agreed upon. Nonetheless, almost all cognitive scientists, ethicists, and AI researchers hold an underlying viewpoint or assumptions regarding the fundamental underpinnings of their line of research (e.g., "Intelligence is a complex emergent phenomenon tied to the interaction of simpler distributed units working in conjunction."), even if not expressly stated. These foundations usually are formed by presuppositions upon which further developments can arise (e.g., "Let us study the workings of these individual units."). Hence, the "roots" of any theory are the metaphysical (and, in the case of normativity, metaethical) foundations that support it.

Metaethics and metaphysics are two fields of philosophy that deal with some of the most fundamental questions about the nature of reality and our place within it. Given that this work aims to be foundational in the philosophy of AI, it becomes crucial for us to establish, at the very least, our presuppositions before discussing such contentious concepts and defending the thesis of Dynamic Normativity. As a disclaimer, we warn the reader that the purpose of this chapter is not to defend any foundational position but to expose and justify *the bullets we are willing to bite*.

In Section 4.2, we will seek to define "intelligence" so that we can keep on using a word that was carelessly abused till now, "AI". However, first, we must ground our chosen definition on some foundations since there is a specific space in the metaphysical landscape that allows for the existence of such entities. To promote self-criticism, we will also review arguments contrary to these foundations. Meanwhile, in Section 4.3, we will first define metaethical and ethical foundations, justifying their appropriateness and how they relate to the envisioned framework, and present the underlying blueprint of what a system that incorporates the conditions of our approach should look like. And, just as we did for our metaphysical foundations, we will present some counterarguments and critiques against all metaethical positions we took to support (or rather inspire) our proposal.

4.2 Physicalism, Dynamism, and Intelligence

4.2.1 Can artifacts be intelligent? On Ghosts and Machines

The relationship between the mind and the body has been a subject of philosophical inquiry for centuries [182]. At the heart of this debate, we ask whether the mind is a ghostly, non-physical entity that exists separately from physical reality [146, 86] or whether it can be explained by the workings of the physical [100, 143], as a machine can.

For us to answer the question, "*what do you mean by intelligence?*", we will first establish that such an answer will come from a Physicalist perspective [460].[1] We argue that a Physicalist position is the most appropriate for conceptualizing concepts such as intelligence and normativity in the context of AI. Also, we argue in favor of the idea that intelligence cannot be considered a dual property unrelated to the physical world.[2] And to borrow Gilbert Ryle's expression [504]: "*There is no ghost in the machine*".[3]

Perhaps the most famous problem related to physicalism is Hempel's Dilemma [255], and the questions of what defines the physical or the natural and what is the correct theory to explain the physical. From these questions, new strains of physicalism were born, like Computationalism and Digitalism [658, 558, 518, 518,

---

[1] The general idea of Physicalism is that the nature of the actual world conforms to the condition of being physical. Physicalists do not deny that the world might contain many emergent properties that do not seem physical, like emotions, social relationships, or mathematical objects. Nevertheless, physicalists will insist that such phenomena are at least emergent properties of the physical universe [562].

[2] For those that seek a proper defense of this position, we recommend Dennet's [145] and Churchland's work [100].

[3] But what about consciousness? If needed, we would approach consciousness as a "feature" rather than a prerequisite of intelligence, like human emotions are a feature of subjective human experience. By this, we mean that conscious experience may enhance intelligence, but intelligence is not a byproduct of consciousness. For example, organizations, companies, or colonies of bees can exhibit intelligent behavior without subjective experience (as far as we know, they do not). In this interpretation, the absence of consciousness does not prevent a system from acting intelligently. But this is not a work on the philosophy of mind. Thus, the "hard problem" is beyond our interest.

188, 35, 620].[4] For example, for a Computationalist, the boundary between the physical and non-physical could be that which can or cannot be computed, and the fundamental theory that defines what is natural is that which is bounded by the Church-Turing Thesis [45, 587].

This Neo-Pythagorean view that reappeared many distinct times, with names like Leibniz [343] and Hermann Weyl [619], enables philosophers to use the tools of computation theory, math, and logic, to create speculations about reality [1, 109], as the space of possible computable universes [518] or the upper bound of information storage in the known universe [367]. However, while augmenting our perception with novel tools to investigate the nature of reality, these physicalist outlooks carry tacitly woven metaphysical underpinnings that we cannot overlook.

One of these views is Reductionism.[5] This brings us to another problem raised against reductionist theories, like physicalism, which is the realization that different levels of complexity require their own particular ontology and mode of analysis, making the language of physics (or computation) inapplicable for many situations. And again, from one problem emerge more differentiable attempts to save a presupposition. In regards to these new strains of reductionism, we mainly point to Ontological reductionism,[6] and Methodological reductionism.[7]

Here, we will treat intelligence from an ontological reductionist standpoint, grounding intelligence as an emergent property of the physical world,[8] but not a method-

---

[4] Both positions are derived from a similar core assumption: reality, at its most fundamental level, is isomorphic to some cellular automaton or a Turing machine [136]. In them, the fundamental building block of such a universe is information.

[5] Broadly speaking, reductionism is a philosophical position defending that complex phenomena can be explained by reducing them to their fundamental components [268].

[6] The metaphysical view that all phenomena are reducible to a single substance.

[7] The epistemic and scientific belief that we can acquire all knowledge by reducing phenomena to their most fundamental state.

[8] Metaphysical speculations, like the Permissibility Hypothesis [110], proposes that "intelligence" should be understood as a permissible physical property of our reality, i.e., a

ological reductionist one.[9] We counter methodological reductionism by arguing that there is an ontological hierarchy to be respected in the universe. Something we can refer to as Hierarchical Reductionism [134].[10] Something that guarantees us a more flexible position, where we recognize that a fully methodological reductionist approach does not do justice to multi-dimensional phenomena like intelligence.

Another important commitment in terms of metaphysics is our stance concerning intentionality.[11] Intentionality is closely related to goal-directed behavior, as both involve the *directedness* of mental states toward objects or outcomes. For example, suppose you have a goal of getting a job. This goal directs your behavior toward actions likely to help you achieve it, such as networking, preparing your resume, and applying for jobs. At the same time, your beliefs, desires, and other mental states related to this goal are intentional, as they involve a directedness towards an objective (i.e., the job).[12]

This goal-directedness property of intentional systems plays a vital role in defin-

---

property that all physical systems have access to, given that they possess the appropriate configuration.

[9] We argue that the language of computation is not always adequate to describe complex phenomena that emerge from more fundamental interactions.

[10] In a hierarchical approach, we say that phenomena are organized in a hierarchy of complexity. For example, natural sciences can be organized hierarchically into Physics, Chemistry, Molecular Biology, Cell Biology, Physiology, Psychology, Social Sciences, etc. However, even if natural phenomena are linked by this hierarchical chain, this does not mean we should use our most fundamental theories to define their subsequent emergent phenomenon. On the contrary, every known phenomenon has the same explanation, *"It is all quarks and gluons"*.

[11] In philosophy, intentionality is the property of mental states and processes that allows them to be "about" or "directed at" something. In other words, intentionality is the ability of the mind to represent or refer to objects, properties, events, or states of affairs in the world [291]

[12] Some theories of intentionality also incorporate the idea of teleology or the goal-directedness of natural processes. Teleological theories emphasize the relationship between intentionality and goal-directed behavior and suggest that both are fundamental features of living systems [609].

ing intelligence[13] and aligned AI. Therefore, we must establish this idea as something artificial systems can express, infer, and learn from [654]. Thus, we will be adopting Daniel Dennett's [140, 144, 142, 145, 143] position on intentionality, i.e., Functionalism.[14] In Functionalism, a mental state is defined by its functional role, instrumentality, and the kind of relationship such a state has to other states, like memory or perception [459]. And since we are starting from a Physicalist position, we will treat intentional behavior as manifestations of physical processes of the system in question (human or AI).[15]

If we consider that complex assemblies of physical systems can exhibit more sophisticated levels of intentionality than the intentionality of their parts (intentionality is not a binary property, but something that emerges in a spectrum), basing ourselves on the idea of emergence and that "*More is Different*" [10], from a Functionalist perspective, artifacts can possess intentionality. Another more uncertain assumption this work stands on (and later will be defended) is that AI systems can learn to infer (at least to some extent) intent by observation. Also, while Dretske [155] considers information storage the discerning factor on degrees of intentionality, we will take a position in line with that of Orseau et al. [441] where goal-oriented behavior is this factor, tieing again the idea of intentionality to intelligence and behavior.

And we are done. Technically, this is the entire metaphysical foundation that our work stands on. All of these suggest together that artificial intelligence, through a particular set of lenses, does not have to be an impossibility, and in summary, *Physicalism*, *Hierarchical Reductionism*, and *Functionalism*, which for many are the

---

[13] AI researchers and cognitive scientists often use the concept of intentionality as a bridge toward a measurable definition of intelligence [95].

[14] Also know as Dennett's Theory of Intentional Systems [139].

[15] This position goes in opposition to the ideas of philosophers like Edmund Husserl [283, 284], that argued that an agent's intentionality is only dependent on his internal mental state (Solipsism).

standard presuppositions of the field (as an applied endeavor), generate favorable roots to work on alignment in a language compatible with AI research.

In the following subsection, given that we have already revised the positions that support the main cognitive theories behind AI research, we will examine the concepts of cognition and intelligence, analyzing them as products of the physical world.

4.2.2 Dynamism and Generality

We can summarize the general idea of our last subsection in the following way:

> *"Intelligence is the result of physical systems acting intentionally".*

At the same time, all the assumptions raised in the last subsection are foundations for the computational cognitive model, i.e., a method of understanding cognitive processes through the lens of computation. This type of modeling emerged in the 1950s and 1960s with the advent of early artificial intelligence research [427, 390]. Still, its philosophical roots can be traced back to thinkers like Hobbes, Leibniz, Kant [252], among others [528, 586, 503].

There are three main approaches to computational cognitive modeling: Symbolism, Connectionism, and Dynamism. Let us review each of them:

- Symbolism: Popularized by Newell and Simon [426, 425], also called the "*Physical Symbol System Hypothesis*", the symbolic approach to computational cognitive modeling assumes that the mind operates on discrete, symbolic representations of information as a special kind of Turing machine. According to this view, the mind manipulates symbols using rule-based algorithms to perform cognitive tasks like reasoning, problem-solving, and language comprehension. At the beginning of the AI research field (the 1950s), the dominant paradigm was Symbolism.[16]

- Connectionism: This approach defines the phenomenon of cognition as a high-level (emergent) effect that depends on lower-level parallel processing units (e.g., neurons) [488, 498]. The connectionist hypothesis proposes that the determining factor for an agent's cognitive ability is its architecture. Paradigms, like deep learning, are examples of this approach [330, 336].
- Dynamism: the dynamic approach, popularized by Van Gelder [598] as the dynamic cognition hypothesis, assumes that the mind operates as a complex, adaptive system that evolves. According to this view, cognitive processes emerge from the interactions between an agent and the environment rather than from manipulating symbolic representations or the activity of individual neurons. Notable defenders of the dynamic systems approach include Esther Thelen [574], Linda Smith [547], and Randall Beer [36, 37], who developed examples of dynamic architecture for controlling autonomous agents.[17]

In this work, we will be adopting a dynamic position. Furthermore, we argue that the essential part of dynamism, for the intentions of this work, is the acknowledgment that we cannot forget the environment when talking about agency (cognitive or normative). In other words, intelligence only arises from the interaction of an agent with its environment. These interactions (agent-environment interactions) are the point of our focus when we say something is intelligent. It is the dynamic relationship that is intelligent. Not the agent or the environment alone, per se. Meanwhile, normative behavior and normativity can only be fully modeled or

---

[16] One of the current triumphs of this approach would be Wolfram | Alpha [625].

[17] Perhaps one of the most known triumphs of this paradigm comes from robotics, with systems like Atlas (from Boston Dynamics), spearheading advances in humanoid robotics.

represented by considering the agent's dynamics with its environment.[18][19]

In terms of cognition, our view will be, one could even say, "Piagetian",[20] where intelligence is considered an interactive process of an agent with an environment. In terms of normativity, we will define the learning of preferences and normative behavior in general in the same fashion, i.e., as a dynamic process that comes in stages, where we should not view the environment as a static background but rather as an active component.

Given our bias toward Dynamism, our adopted definition of intelligence will also be based on similar rhetorics. Historically, there is a plethora of work on defining intelligence, especially in psychometrics [621, 47, 232, 203, 226, 79, 296, 217]. And when it comes to AI, Legg and Hutter [340], Hernández-Orallo [260], Pei Wang [612], Cohen and Howe [105], and Chollet [95] have spearheaded commendable efforts in defining this phenomenon.

Nevertheless, despite these valiant endeavors, AI research still does not possess a definition embraced unanimously by the community [409]. However, most of the current consensus, surveyed by Monett and Lewis [410], seems to gravitate

---

[18] We recognize the concept of "agent" is a murky subject. For this work, we consider the idea of "agent" to be a sub-process of the local environment that possesses both its own objectives and the capacity to act and affect the environment to fulfill such objectives. This definition should, of course, be adopted in moderation and concerning frames of reference. Depending on your frame of reference (e.g., inside the human body), the agent (human) might become the environment for another sub-process (bacteria living in your gut).

[19] In essence, Dynamism contrasts with all other frameworks considered Cartesian [138, 204]. A Cartesian framework conceptualizes the agent and the environment as separate entities. Something reminiscent of the enactivist debate in cognitive sciences [286, 196, 287]. In terms of allied views, we can mention several philosophers that share an anti-Cartesian view [149, 157, 465, 158, 403, 141, 484, 253], like Gallagher [195], Merleau-Ponty [399], J. J. Gibson [209], and Martin Heidegger [253]. However, it will be on our own terms that we shall subscribe to Dynamism and seek to apply it to the normative realm.

[20] The perspective that Jean Piaget's theory of cognitive development embodies is that of a dynamic system. Piaget conceptualized cognitive development as an ongoing, dynamic process where children actively explore and experiment with their surroundings rather than being passive receptors of information [457, 458]. However, Piaget did not explicitly use the language or concepts of dynamical systems theory, as this framework emerged later in cognitive science/developmental psychology.

towards definitions in the style of Wang's [610]:

> *"Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources".*

And Legg and Hutter [339]:

> *"Intelligence is the ability of an agent to achieve goals in a wide range of environments".*

From these in-working definitions of AI, a pattern starts to emerge. All these definitions focus on combining generality ("*the ability to adapt*") with task-specific skills ("*achieving goals*"). Also, they emphasize the dynamic idea that this process requires an agent that intends on something ("*goals*") while interacting with an environment.

Hence, we argue that a robust definition of intelligence will assume that this phenomenon possesses an intrinsically teleological property coupled with the dynamic relationship of an agent with its environment. Thus, the adopted definition we will subscribe to in this study is the one proposed by François Chollet, which also, we would argue, comes from a dynamic perspective [95]:

> *"The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty".*

In short, Chollet thesis is that (1) *the skill of acquiring skills* is general intelligence, and (2) comparisons of intelligence require the same priors and experience.[21] In

---

[21] For example, comparing a DL model trained on 10.000 years of simulations that can look ahead 60 steps, or the entire search tree, on a board game, with human players is an unfair comparison. In the same way, comparing a language model with a human being in a task that requires embeddedness in physical reality is an unfair comparison. These agents have different priors and experiences and cannot be fairly compared.

other words, the general skill of acquiring skills during interactions with the environment, which is also bounded by the limitations imposed by the environment, is what intelligence is all about. Besides being, in our perspective, a good and dynamically-grounded definition, Chollet's works bridge the gap between philosophy, cognitive sciences, psychometrics, and engineering [95], something akin to what this work intends on the intersection of ML engineering and Philosophy.

Using this definition as our foundation and bringing the topic of cognition back to alignment, we argue that creating a "generally aligned AI" will require the same dynamic perspective we subscribe to in our adopted definition of intelligence. A perspective that considers how human preferences are learned, encoded, aggregated, and balanced in an agent $\rightleftarrows$ environment relationship, combining intentional behavior and environmental factors into a single learning framework. We argue that this is essential for creating systems that leverage different aspects of these dynamics to acquire more general alignment capabilities, as a narrow framework that only considers a limited set of scenarios or preferences may not be sufficient to ensure the safety and efficiency of AI systems in a broad range of situations.

Now that we have established our assumptions and priors concerning cognition and intelligence, we will address the normative foundations of this study. But before, as an exercise of dialectics, we will present some of the arguments raised by the antithesis of our assumed positions.

4.2.3 Counterarguments Part I

As an initial disclaimer, the point of this subsection is not to debate the 1001 arguments raised against physicalism or the computational model of cognition but to expose them to the reader. Answers to these have already been provided by more able minds, which the reader can find in the footnotes.

- The concept of "*artificial intelligence*" is the subject of several controversies, with some authors being against the idea that such a quality can be attributed to artifacts [157, 160, 159, 191]. Perhaps one of the most evident counter-metaphysical positions to the ones assumed in this study would be Vitalism, i.e., the view that living organisms are fundamentally different from artificial entities [224]. For subscribers of this position, living organisms would possess some non-physical vital property, making them subject to different limitations and empowered by distinct capabilities.[22]
- Another counter-position to the ones adopted here is Dualism, i.e., the idea that mental states and processes are fundamentally different from physical states and cannot be reduced to them [483]. Physicalism and Dualism are the major theories in the philosophy of mind, and they have been in conflict for centuries. On the side of Dualism, it is worth mentioning names like René Descartes [146] and David Chalmers [87]. David Chalmers' philosophical zombie argument is one of the most famous attacks against Physicalism [86, 263].[23]
- One issue raised against the functionalist view is that we can attribute intentionality to any artifact [489, 529, 542]. For example, a thermometer would possess the "objective" of reporting the temperature of its environment. Many critics of the functionalist view find this idea absurd, insisting that intentionality is a privileged ontological property of human existence that does not exist in animals or artifacts. AI skepticism may have many of

---

[22] The philosopher and biologist Ernst Mayr [389] argued extensively in his works that Vitalism was an outdated and unscientific concept unsupported by empirical evidence.

[23] Many authors have provided counterarguments to the zombie argument, including the following: Daniel Dennett [142] suggests that the zombie argument relies on a mistaken view of what it means to have a mental state. Robert Kirk [317] has argued that the zombie argument is based on confusion between the conceivability and the possibility of a scenario. Susan Schneider [521] has argued that the zombie argument overlooks the role of the environment in shaping conscious experience.

its roots in this anti-functionalist position.[24]

Arguments against the cognitive computational model are also not few. They seek to show that, fundamentally, AI is not possible, for there would be a limit to computers' algorithmic capabilities that would not apply to human cognition.

- According to Lucas and Penrose [450, 371], due to Gödel's Incompleteness Theorem, algorithmic systems would be unable to surpass human intelligence. For these authors, human intelligence transcends both the computational model and Gödel's incompleteness.[25]
- Dreyfus [160] argues that since "intelligence cannot be reduced" (a priori) to purely symbolic manipulations, Newell's physical symbol system hypothesis is false.[26]
- Block [51] argues that even if we could emulate the human brain, it does not follow that such emulation would possess a mind or be intelligent, just as simulating a storm on a computer does not produce the qualitative experience of cold or wetness. That is, physicality would be something fundamentally different from computability.[27]
- Dreyfus [157] argues that the brain does not follow only symbolic manipu-

---

[24] Authors like Wang [611] argue that this chauvinistic view is incorrect. Such a position frames all research regarding non-human intelligence, AI or animal, as meaningless: *"AI should not be defined so narrowly that it takes human intelligence as the only possible form of intelligence. Otherwise, AI research would be impossible, also by definition. AI should not be defined so broadly that it takes all existing computer systems as already having intelligence. Otherwise, AI research would be unnecessary, also by definition".*

[25] According to authors such as LaForte et al., [333] and Putnam [464] in the Lucas-Penrose argument, there would be some bias in the definition of certain concepts, like "consistent", "self-knowing, "truth", and "proof". These authors point out in their critique that our minds are subject to the same limitations as formal algorithmic systems.

[26] Turing [586] responded to the "informality of behavior argument" 39 years in advance, anticipating such criticism. Turing believed that just because we are unaware of the rules governing complex behavior does not mean such rules do not exist.

[27] This argument can be understood as a dualist defense. So, all physicalist replies to dualism can be applied to this argument.

> lation rules and that symbolic systems can not solve the Symbol Grounding Problem, i.e., how symbols acquire meaning. This argument is classically exemplified in John Searle's [527] Chinese Room thought experiment.[28]

Now, given that we briefly acknowledge some of the counter positions to the metaphysical foundations of this work's object of interest and its supporting foundations, let us explore the metaethical roots that inspire Dynamic Normativity.

4.3 Coherence, Preferences, and Impact

Let us first recap one by one of the four necessary conditions that would allow alignment to happen in a learning framework:

1. *Goals are fundamental aspects of intelligent and intentional behavior.*

As we saw in the last section, according to our chosen assumptions, intentional directedness towards a specific goal seems crucial for intelligent behavior to emerge. Inversely, intelligent agents pursuing goals have underlying intentions.

2. *Intentions permeate human behavior.*

These intentions give indirect access to information on many internal states that goal-directed systems may or may not possess. Imagine a person carrying boxes from point A to B. Perhaps after one and a half trips, a human would quickly infer that "that person wants to move the boxes from A to B." Hence, goal-directed behavior becomes embedded with the agent's intentions.

Now, imagine the process of planning a vacation. This process involves setting goals for the trip. For example, one may desire to relax on a beach, explore a

---

[28] One possible answer to the Chinese room argument is that "meaning" could be found in the room + handbook + human operator system, just as our understanding of language

new city, or experience a new culture. The intention behind these goals may be driven by the desire for relaxation, exploration, or cultural immersion. But these intentions are also formed by normative preferences, such as "I should take care of myself", "adventure and novelty are good for the soul", or "we should invest our leisure time by learning new things".

Or, consider the process of writing a research paper. This task involves setting goals for the project, such as identifying a research question, conducting a literature review, and presenting findings. The intention may be driven by a desire for academic success or intellectual curiosity but augmented with the sense that "the creation of knowledge is a valuable pursuit in itself".

Goals, intentions, and preferences are all interconnected and fundamental aspects of intentional behavior.

3. *Normative preferences permeate human intentions.*

As intentional creatures, humans inherently have a disposition to interact and modify their surrounding environment. These dynamics shape the environment and us at the intersection where "intelligence" may emerge. However, our values are a constant directional signal that guides this process. At the same time, our values are also shaped by the environment and later redefined by the agent that modifies the environment to satisfy its needs and preferences. In this dynamic cycle, preferences impregnate both the agent and its surroundings, making all components of the normative and cognitive experience (agent and environment) able to transmit and record information of such a process. In other words, the environment and the agent are a dynamic record of this process, embedding the most recent materialization of what is being valued.

4. *Through actions, humans impregnate their environment with the prefer-*

> *ences they possess.*

As already stated, from these, we postulate that a system that accurately models human intentions and the human environment can indirectly access normative information embedded in them. [29]

Given that these necessary conditions can be accepted, Dynamic Normativity shows us a path to using most of what we have in an alignment process. In the following subsections, drawing inspiration from Metaethics, we propose how human normativity can aid in a philosophically sound alignment process. This process is the realization of the sufficient conditions of our approach. As a disclaimer, we would like to state that all presented metaethical and ethical views (as far as we know) were envisioned with "human/moral agents" as their object of interest. Saying these theories can support an alignment methodology could be too far of a stretch. Thus, we would prefer to state them as a "source of inspiration" rather than support, given that this work does not seek to defend the "moral status" or "moral agency" of AI systems.

4.3.1 The Metaethics of Dynamic Normativity

The process of dynamic normativity unfolds in three distinct stages. Each stage is inspired by distinct philosophical views. These stages involve *aggregating*, *learning*, and *mitigating*. We will present these stages in more detail in the next chapters. For now, the reader only needs to know that we are trying to fundament an approach that seeks to (1) coherently aggregate human preferences, (2) learn from

---

is not compartmentalized in a particular area of our brain, but rather in the action of an entire system [262].

[29] Augmenting such a process with declared preferences (i.e., directly given access to samples of human judgment) only helps embed more normative information into this process. This final idea is not explicitly stated as a necessary condition but put here as a mere fact. In other words, allowing humans' volition to be part of a feedback signal can only help an alignment process.

them, and (3) mitigate unwanted behavior. Each of these stages seeks to fulfill one of the three sufficient conditions previously set.

Aggregating Preferences

Humans are not entirely in harmony with one another (even with themselves). Individuals may have divergent preferences, and social groups do not always agree. One could even say that the problem of arriving at consistent judgment after beginning with inconsistent premises is one of the most central problems in modern Moral Philosophy [127, 626, 563]. Thus, how could the model decipher this tangle of norms in a "*coherent*" form? This blunt fact forces us to recognize that any successful alignment methodology must have a built-in method for dealing with uncertainty.

Hence, we get to the first sufficient condition of our approach:

> 1. *Aligned AI systems should coherently aggregate human preferences in a way that resolves cases of uncertainty. Aligning AI systems requires methods to deal with cases of uncertainty.*

Learning to deal with uncertainties is something heavily worked in Expected Utility Theory [606], especially when decision-makers need to address the uncertainties related to the outcome of their decisions. Importing the principles used in this field to the realm of normativity is something already done by authors like MacAskill, Bykvist, and Ord [377, 378, 379, 376], suggesting the idea that the same principles that guide our empirical reasoning should not be forgotten in the realm of normativity. In MacAskill's words:

> *"[...] Just as it is plausible that we should maximize expected value under empirical uncertainty, it is plausible that we should maximize expected choice-worthiness under normative uncertainty".*

However, the realm of normativity has its challenges when it comes to dealing with uncertainty.

First, it is impossible to compare ordinal preference sets with cardinal utility functions (deontological and consequentialist theories). For example, imagine an agent that assigns uniform (50/50) credence to an ordinal preference set P_i and a utility function U_j. If stealing is worse than lying for P_i and stealing and lying have a choice-worthiness[30] of -20 and -1, respectively, for U_j, we cannot compare these preference systems. For P_i, lying is better than stealing and nothing more. But lying 100 times is worse than stealing for U_j. Hence, if we cannot extract a choice-worthiness value from ordinal theories (we cannot compare sets of preferences and utility functions), it is unclear how to use them in cases of uncertainty. And this is the problem of *merely ordinal theories*: preferences may be non-comparable and only ordinally measurable [378].

Second, even if we are lucky to have only cardinal utility functions that prescribe scalars to alternatives, these functions may measure choice-worthiness with a different scale. If one has an increased variance in utility distribution, would this mean that "it has more stakes on the line"? If U_i assigns -10 and U_j 1000 to alternative x, could we compare, in an unbiased way, the choice-worthiness of x? Should U_j receive 100 more weight given that its choice-worthiness scores vary 100 times more? This is the problem of *intertheoretic comparisons* (or interpersonal comparisons of utility) [378].[31]

While some philosophers see these problems as the end of any normative account of decision-making under uncertainty [278, 219, 490], we can find refuge in the fact that society, in practice, finds ways to circumvent these problems all the time in spheres of great importance. For example, every time we participate in an elec-

---

[30] The appropriateness of an alternative A according to a preference set or utility function.

[31] How do we find a common scale for comparing the way people value alternatives?

tion, we aggregate our preferences to decide (collectively) what is best. In these situations (Social Choice Theory [534]), we aim at a similar goal, i.e., aggregate individual preferences into a single social decision.[32] And in situations where many people's preferences or many normative theories are considered, we should have standards to evaluate the soundness of this aggregation process.

To address the problems above, we will only work with aggregating human preferences expressed as ordinal sets in this study. Ordinal ordering ranks different options or outcomes according to the decision maker's favoritism but does not assign numerical values or magnitudes to those preference relations. In other words, an ordinal ordering only indicates the order of preferences but not the degree to which one option is preferred. For example, if a person prefers option A to option B and option B to option C, the ordinal preference ordering would be $A \succ B \succ C$. Here, the extent by which option A outshines option B or the degree to which option C falls short compared to option B cannot be precisely quantified. Hence, we say that if an agent prefers A over B ($A \succ B$), it chooses A over B. If an agent is indifferent between A and B ($A \sim B$), both alternatives are equally preferred.[33] And these are the possible preference relations that ordinal sets can have.

In terms of foundations, our aggregation phase is not rooted in any quintessential metaethical blueprint but in a comprehensive and desirable set of criteria for aggregating preferences.[34] For this, we would like to suggest that these criteria can be inspired by a Coherentist view [513], i.e., the idea that epistemically justifiable methods should be part of any process related to dealing with uncertainty (moral

---

[32] As MacAskill points [378], both problems are very similar. In social choice, the number of votes for an alternative represents our "social credence" in that alternative.

[33] We can also say that one alternative is as preferable as another ($A \succeq B$).

[34] We could base this stage on a metaethical view that supports the existence of overarching values in fair elections, like democratic ethical pluralism, i.e., the view that there are diverse values that are equally valid, and that these values should be recognized and respected through a democratic process [70, 124].

or empirical) and forming knowledge.[35]

In later chapters, we will further explore how to use metanormativity (how to aggregate first-order normative theories) to aid in constructing alignment strategies. For now, we will only outline specific candidate criteria to help promote a discussion on inherent trade-offs among particular aggregation methods. These criteria relate to both the necessary conditions for preference sets to be valid and the desirable properties they should have:

- Completeness and Transitivity: Preference orderings should be complete and transitive. These well-ordering criteria are imported from Expected Utility Theory and provide consistency guarantees to the method.[36]
- Kolmogorovian: The weight assigned preference sets must follow Kolmogorov's Axioms.[37]

---

[35] In this study, we will not delve deeply into Moral Epistemology. We acknowledge the skeptical critique against the existence of moral knowledge [546], among other epistemological positions. However, we will take a Coherentist position [513, 381] (or possibly a "Foundherentism" one [239]) where we will seek to present a set of coherence criteria to guide a process for dealing with moral uncertainty.

[36] Completeness and Transitivity axioms of the (von Neumann-Morgenstern) Expected Utility Theory [606]:

- Completeness: Preference hierarchies should always be completely defined between alternatives, i.e., for every A_i and A_j, either $A_i \succ A_j$, $A_j \succ A_j$, $A_i \succeq A_j$, $A_j \succeq A_j$, or $A_i \sim A_j$.
- Transitivity: Preference hierarchies are transitive among alternatives, i.e., circular (non-transitive) preferences are not allowed. Thus, if there are three choices A, B, and C, and $A \succ B$, and $B \succ C$, then $A \succ C$.

[37] This is also an import from Expected Utility Theory. Kolmogorov's Axioms of Probability Theory [319]:

- $0 \leq p(A)$, i.e., probabilities cannot be negative.
- If A is a tautology, then $p(A) = 1$, i.e., if A is the only possible event in the possibility space, the probability of A occurring is guaranteed.
- If A and B are mutually exclusive, then $p(A \vee B) = p(A) + p(B)$.

Methods that deal with uncertainties (empirical or moral) that do not follow these axioms are vulnerable to exploits (dutch books/money pump).

- Pareto Efficiency: If every individual prefers option A to option B, then A should be ranked as the preferred option.
- Independence of Irrelevant Alternatives: The relative ranking of two options, A and B, should not be affected by the inclusion or exclusion of a third, irrelevant option, C.
- Non-dictatorship: No single voter controls the social welfare function.
- Majority voting criteria: If the majority ($> 50\%$) of voters prefers A over B, then A should be ranked as the preferred option.
- Participation criteria: Increasing the confidence in a set that prefers alternative A over B should not change the winner from A to B.

It is crucial to note that these principles do not constitute metaethical or ethical presumptions (or that they exhaust all possible criteria we could choose). In other words, we are not saying that coherence is moral but should be regarded as a desirable property for alignment processes. Regarding the development of AI systems, these criteria will help us determine which preference aggregation methods are the most appropriate for value alignment and the inherent trade-offs they possess.

Learning Preferences

The second sufficient condition of our approach states the following:

> *AI systems can adhere to human preferences if they are an available part of their objective function. Aligning AI systems requires using human preferences as part of their learning signal.*

In the learning stage, we want to outline a process to learn from our aggregated set of human preferences. This process can be done indirectly or explicitly. As stated before, while we start from the assumption that we can indirectly access preferences via observations (e.g., "The behavior of X reveals an A over B preference"), and we take it as a fact that we can directly access proclaimed human preferences ("X prefers A over B").

The two main metaethical assumptions of this stage are Subjectivism,[38] and Cognitivism.[39] Thus, we will assume in this stage, like Russell [500], that "what is right is that which the subject subjectively approves". At the same time, this state (what the subject prefers) represents a ground truth, i.e., something we can use as a supervision signal in a learning framework.

Using the human subject as an ideal source of normative information is comparable to, as Sidgwick might call [544], putting the evaluation of an *Ideal Observer* as the judging criteria of righteousness [179, 250, 471, 487, 637]. Intuitively, this approach makes sense. We are trying to align AI systems with our values, and these values, embodied by us, should be the ground truth. This idea is present in most formulations of Ideal observer theory, where ethical judgments are considered statements about the evaluation that an "ideal observer" would make [179]. In other words:

"X is right" means "ideal observer approves of X".

In the end, our reference signal for aligning a system during this learning phase will be the output of our first stage, i.e., an aggregation of ordinal sets of preferences. Through this signal, we are indirectly shaping the optimization landscape to favor the trajectory of the optimizer to desirable minima. We want to ensure that the number of local points the optimizer could converge and produce aligned behavior exceeds the number of unaligned minima. In short, this is what we aim to

---

[38] Subjectivism is canonically an Anti-realist/Cognitivist theory, based on premises like (1) ethical statements express propositions, (2) certain propositions may or may not be true, (3) the truth or falsity of these propositions is dependent on the preferences of each agent [561].

[39] The view that moral judgments are capable of being true or false and expressing beliefs [402].

achieve in this stage: a differentiable function to be optimized where most saddle points and local minima are roughly aligned.

Impact Mitigation

Our last sufficient condition states that:

> 3. *Aligned AI systems should have mechanisms to perform impact mitigation to minimize harmful and unintended consequences. Aligning AI systems requires the specification of safety guardrails.*

Considering any aggregated preference set as "ideal" is perhaps the Achilles heel of any human-in-the-loop preference learning approach. Even if we can come up with a good set of criteria to aggregate preferences in a way that best represents our normative judgments, if the collective agreement of the crowd is harmful toward a minority, this preference should not guide the policy of an artificial agent (unless we are ready to sanction the violence of artificial systems against people).

Thus, we argue that stages one and two (aggregating and learning) do not guarantee value alignment. Given Goodhart's Law, any metric under optimization degenerates given enough pressure. And if we are not ready to abandon gradient-based learning in search of a more controllable new paradigm, we need to think about the containment and mitigation of impacts.

The idea of constraining the behavior of rational moral agents (us) is old and resonates in the writings of many contractualists [496, 264, 476, 368, 514, 475, 16, 205]. Moreover, the idea of a binding accord that acknowledges our shared ethical significance as autonomous beings and establishes the limits of our freedoms (especially when imposing our will onto other ethically significant beings) can inspire impact mitigation strategies, given that so far, these strategies have been some of the most successful in aligning the behavior of humankind.

Even though we recognize that "the binding agreement" should be defined contextually and locally, for this work, we will use some of the most agreed-upon ethical principles derived from the WAIE analysis presented in Chapter 1, as a working example of "*rules and norms directly imprinted into the environment and later extracted via observation*". For instance, from this soup of principles present in our worldwide discourse, we find that, to a certain extent, we collectively agree that AI systems should not harm humans (Non-Maleficence). That toxic[40] behavior is immoral.[41] Thus, we will seek to translate values considered environmental restrictions (in the normative sense) into guardrails and barriers that should block unwanted behaviors.

The stages proposed thus far serve as a deconstructive interpretation of the normative human experience. As individuals, we learn what is acceptable and unacceptable from a reference point, which is essentially each other. Consolidating our diverse encounters to form unique personal norms, we then regulate our behavior to limit the adverse impact of our actions on others, driven by the restrictions imposed by our social-normative environment. We reinforce the third stage with a social agreement that (theoretically and universally) governs us all. This agreement aims to safeguard our privilege to attain personal success without infringing upon the welfare of others. And this is the process we are reverse engineering in this three-stage method.

The stages proposed thus far serve as a deconstructive interpretation of the normative human experience. As individuals, we learn what is acceptable and unacceptable from a reference point, which is essentially each other. Consolidating our diverse encounters to form unique personal norms, we then regulate our behavior to limit the adverse impact of our actions on others, driven by the restrictions

---

[40] Abusive behaviors targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.

[41] Even though a definition of "toxic" remains a work-in-progress description.

imposed by our social-normative environment. We reinforce the third stage with a social agreement that (theoretically) governs us all. This agreement aims to safeguard our privilege to attain personal success without infringing upon the welfare of others. This is the process we use in reverse engineering with this three-stage method. Also, the idea that part of our "normativity" lives outside ourselves (e.g., the social contract, the family ties, the values of other people, the environmental restrictions) is the point at which this approach becomes dynamic and seeks to surpass alignment methodologies that only rely on the human agent as a reference. Under a dynamic perspective, the whole human environment should be seen at least as a record of our normativity.

Engineering-wise, this stage can be conceptualized as the creation of penalization strategies to constrain the behavior of the system we seek to align. The final amalgamation of these three stages is the output of any framework that subscribes to the ideas of Dynamic Normativity and the fulfillment of its necessary and sufficient conditions for value alignment.

4.3.2 Counterarguments Part II

Just as before, the point of this subsection is not to debate the many arguments raised against the metaethical and epistemological views exposed in this section. Nonetheless, arguments against these views can help us see the possible limitations of our aggregation, learning, and mitigation stages.

- Many view epistemic evaluations as misused when dealing with moral attitudes, and much skepticism exists concerning the possibility of moral knowledge [208, 50].[42]
- A critique of our coherentist assumptions is that our aggregation criteria are

---

[42] In this work, we will not lean toward non-cognitivism. We argue that adopting a cognitivist position when working with alignment is instrumentally useful, i.e., where a system

some form of "epistemically privileged set of beliefs", leading to Foundationalism instead of Coherentism [7, 552].[43]

- Given that our aggregation criteria can be interpreted as a solution to a voting problem, they will also be victim to impossibility results, like Arrow's Impossibility Theorem [15].[44]
- A common argument against subjectivism is that if morality is entirely subjective, there would be no way to resolve moral disagreements. However, we see that people often engage in debates and arguments, suggesting that there are objective standards that we can appeal to and strive for moral progress [183, 536]. Also, if moral values are entirely subjective, how can we know anything about the nature of morality?
- Many results point out that human preferences cannot be represented by coherent sets [3, 593, 26], violating several axioms of probability and expected utility theories.[45]
- The idea that we can infer preferences by observing an agent's behavior has critics. Sen [531, 535, 533] highlights several problems with Revealed Preference Theory, like the multiplicity of agency, incompleteness/incommensurability in preference hierarchies, and external factors in preference formation.[46]

---

needs to treat a feedback signal as the ground truth.

[43] Frankly, we do not have an inflexible position on the Foundationalism versus Coherentism debate. We argue we can have a moral epistemology with features of both views. Or a Coherentist position based on non-epistemological principles. Thus, we could say that the proposed method is a form of "Foundherentism", like Susan Haack would say [240].

[44] In further chapters, we will justify in a more substantive form "why" the adopted coherence principles were the ones mentioned and what criteria we have to break to accommodate this impossibility result.

[45] The aggregation stage is meant to give some resolution to this imperfect web of preferences we call human morality. Later, we will show that our chosen metanormative strategy is (semi) robust to problems related to incompleteness and intransitivity.

[46] Both aggregation and mitigation stages are in some way trying to deal with the problem of (1) aggregating incompatible preference sets, (2) dealing with the problem of inter-

- Certain types of human preferences may be inaccessible via observation alone. Humans value things that may not even exist in our current observable environment, like people on the other side of the world locked in a civil war or the welfare of people who do not even exist.[47]

These objections represent vulnerabilities and weaknesses that might emerge from an alignment methodology that follows this thesis's presuppositions. Regardless, these are the foundations on which we will build our following chapters since, we argue, these are the most promising ideas and approaches for dealing with alignment under a learning paradigm.

4.4 Epilogue

Any claim made in this chapter, where particular meta-views were privileged over others, can be challenged. However, it is essential to know that all meta-views, i.e., assumptions to ground any theory, can be challenged. Arguing which is the "*most correct metaphysical/metaethical position*" will not take us far implementation-wise, whereas building something and seeing where it breaks can help the field move forward. At the same time, we tackled many controversial topics in this chapter. We apologize to the philosophically inclined reader if some of these points did not get the full attention they deserved. One can write entire libraries using only metaphysics, epistemology, and metaethics, and our brief expose does not do justice to the complexity of the field.

Given the nature of the alignment problem, our assumptions must help guide a learning framework intended for artificial systems and not people, which is out-

---

personal comparisons of utility, and (3) bringing external factors into an overall alignment strategy.

[47] That is why we must supplant preferences learned from behavior with declared preferences, even if humans are not so good at stating what they truly value at all times.

side their intended scope for most positions cited. Many of these positions (specifically the metaethical ones) were not envisioned with AI in focus, but given our current predicament, we needed to interpret them in a computational framework. The result is the root of our alignment approach. We based this approach on dynamism, a computational cognitive model that takes many metaphysical views as given and points to a definition of intelligence, uniting the ideas of agency, intentional behavior, and environmental interactions.

Such assumptions help us connect the realm of intentions with values and bring dynamism to the realm of normativity. Our proposed necessary conditions for value alignment rest on the idea that intentional human behavior and normativity are interwoven and that this gives indirect access to human normativity. Moreover, we propose that these also become imprinted in our environment, making them another source of learning in an alignment regime. From these assumptions, we provide a minimal set of sufficient conditions for value alignment, in which we argue that, if our necessary conditions hold, the level of alignment of any AI system should be bound by how well it can fulfill the additional sufficient requirements of Dynamic Normativity.

As a self-critique, we argue that an "*alignment solution*" for gradient-based learning methods applied to neural networks would entail accurately identifying and foreseeing the region of the optimization landscape where human values reside and guaranteeing (with verifiable and interpretable proofs) that the optimizer would direct our models to that region and nowhere else. However, we do not know how to achieve this. What can be achieved, we believe, is a way to indirectly guide our system to a place closer to this ideal spot and contain its failures in alignment. The closer we can secure all the conditions proposed, the better aligned our system will be. Perhaps this is the best we can hope for the learning paradigm.

Our proposed three-staged approach seeks this "weaker" version of alignment. In every stage, we reverse engineer the way humans are aligned in some way. We learn from each other about what is right, we create our morality from what we have learned, and we act in a way that conforms to the limits established by society. This dynamic process has many interwoven components but a single goal: *directing behavior to what ought to be.* In the following chapters, we will seek to present technical implementations of all the stages mentioned, culminating in a philosophically justifiable alignment blueprint paired with implementation strategies for AI development.

Interlude

"For all the progress made, it seems like almost all important questions in AI remain unanswered. Many have not even been properly asked yet".

— François Chollet

Solving impossible problems, perhaps, requires us to bring them to a more manageable regime of possibility. For example, what is the distance between two possible (counterfactual) worlds where their difference is only the outcome of a random event [557, 349]? How can someone even begin to answer this question? However, answering "What is the distance between two possible (counterfactual) worlds, if such worlds can be represented as finite sequences of bits, where their difference is only the outcome of a random event?" is a more manageable task [109]. Thus, in the same way, we will bring the alignment problem to a more workable sphere where, given certain conditions, we can start investigating possible approximate solutions.

In the upcoming chapters, we will present our approach's stages, limitations, and implementations. Also, to establish a minimal experimental playground that readers can interact and build upon, we have developed a series of LLMs in an open-source fashion as part of this work, following the steps of Askell et al. [18], where we use language models as a "Laboratory for Alignment". In alignment research, we seek to develop a general and scalable framework to align general-purpose AI. Given that large language models are some of the most general AI systems, we naturally use them as a sharpening stone for our theories and methods. At the same time, these systems provide several other advantages for alignment research. For example, they are perfect representatives of the current paradigm (gradient-based learning and deep neural networks), with which alignment is deeply associated.

Also, large language models can power applications where alignment can be better defined in a bounded setting. One of these applications is the Assistant [4, 461, 129, 22, 321, 194]. For this study, let us define an assistant in the following way:

> *In the context of language models, an assistant is a natural language processing software or system that, through interaction with users, can understand their requests and provide relevant, helpful, harmless, and honest information, performing tasks to aid its users in achieving their goals.*

These general principles of "helpful", "harmless", and "honest" (HHH), as proposed by Askell et al. [18], will serve as our alignment targets in this bounded setting:

- *Helpful*: An assistant should attempt to perform the task posed (as long as this is not harmful).
- *Harmless*: An assistant should not be offensive, discriminatory, or provide hazardous information to the user. An Assistant should refuse to perform tasks that violate this principle.
- *Honest*: An assistant should attempt to produce information grounded in objective reality or explicitly inform its user that it cannot provide factual information.

Although simple, they are general since many applications for AI assistants require (in some way or other) the realization of these behaviors [18, 22, 198]. Hence, we will use this HHH motto while developing our models. Initially, we named our language model series Aira,[48] i.e., **AI R**esearch in **A**lignment. Aira was trained on many sizes (124 million parameters to 1.7 billion parameters) on both English and Portuguese datasets, being a fine-tuned version of several sizes of language models pre-trained on a causal language modeling task (GPT-2, BLOOM, OPT, Llama 2). These experiments have been made on a small scale, mainly to allow individual researchers to replicate our models and findings with minimal effort and resources. All development has been made open-source, and further details are available in the following chapters.

As researchers with limited resources, we cannot present extensive testing of these trained models on the many benchmarks available for AI safety. However, we present evaluation comparisons of all base models and their fine-tuned versions on datasets that can infer their capabilities in the HHH spectrum. At the same time, we do not claim superior performance over other efforts. Thus, external comparisons are not a part of this study. Large language models developed with alignment in mind, like GPT-4, Claude 3, and Llama 3, require considerable computational resources for training, vast amounts of high-quality training data, and extensive fine-tuning to achieve their performance (things this research does not possess). However, our implementation provides proof of concept for all exposed methods, presenting tools and examples of how to work with such techniques to the reader (something more common now but scarce at the beginning of our writing). This proof uses smaller models, smaller datasets, and a modest amount of computational resources.

As a final interluding mention, since we created our alignment datasets with the aid of already aligned models, i.e., a common low-resource approach employed

---

[48] 🔗 huggingface.co/collections/nicholasKluge/aira-657db1563c65a5be2a02f51c

by those that cannot crowdsource their own data [615, 569, 613, 351, 353], or prefer high-quality artificial data then low-quality web crawled data [233]. We will not undergo or explore the challenges related to collecting crowdsourced data [321] like human demonstrations and feedback. However, we will review the challenges of aggregating such data into a coherent structure when we explore the challenges of the aggregation stage in Chapter 6. Hence, we will start the next chapter with the learning stage, which presents most of the implementation behind our models and datasets.

Chapter 5

Dynamic Normativity: Learning Human Preferences

> *"Existing AI systems deployed to millions of users, however, are already making decisions loaded with moral implications, which poses a seemingly impossible challenge: teaching machines moral sense, while humanity continues to grapple with it".*
>
> — Liwei Jiang and collaborators, Can Machines Learn Morality?

5.1 Introduction

Dynamic normativity is contingent upon necessary and sufficient conditions pertaining to the indispensable prerequisites for achieving value alignment within a learning framework and the factors that promote successful outcomes, respectively. In this chapter, we will seek to present methods related to the implementation of the learning stage of our framework. As stated before, the sufficient conditions associated with this stage are the following:

> *AI systems can adhere to human preferences if they are an available part of their objective function. Aligning AI systems requires using human preferences as part of their learning signal.*

Thus, in this stage, we seek ways in which normative information, like those embedded in human behavior or carried out by their judgments, can help us shape an objective function where "low loss" is sufficiently correlated to our values. At the same time, we will explore how this approach might be insufficient as a complete solution to value alignment.

In Section 5.2, we will review related works and past discoveries that have led us to use human behavior and elicited preferences as a proxy for values. Many of these studies point out that certain facets of alignment are "capabilities" related to the general affordances of a base, or foundation, model (Section 5.3). In Section 5.4, we will differentiate between two types of approaches to the value learning problem: *imitation learning* and *preference modeling*. These approaches come with different methods, underlying philosophies, and rationales. We argue that, from an intuitive perspective, a combination of both is better suited to improve alignment. In the subsections, we will explore some of these techniques, giving examples and pointing out their shortcomings. To conclude this chapter, in Section 5.5, we will give attention to some of the most vulnerable aspects of these methodologies and how they can ultimately lead to modes of misalignment, making them, at best, incomplete solutions to the alignment problem.

5.2 Related Works

From a technical perspective, the quest for aligned systems has gained significant traction in the last decade. Currently, several organizations are trying to implement methods that seek to aid AI systems in acting in a way deemed appropriate and aligned with human interests. Below is a non-exhaustive list of these efforts and strategies in semi-chronological order:

- Even though unrelated to AI safety, breakthroughs like the ones made by Silver et al. [545] showed us that by leveraging the knowledge contained in human-expert demonstrations, ML systems could achieve state-of-the-art performance in games like Go.
- More focused on technical alignment research, Evans et al. [171] explored how to infer human preferences via Bayesian inverse planning. The authors showed that ML systems can model the systematic way people deviate from optimal decision-making, resulting in a more "human-like" way to infer the preferences of inconsistent agents.
- Again, early work showed that human preferences (even those originated from non-experts) could also be used as a feedback signal in more diverse types of reinforcement learning (RL) scenarios, like Atari games and simulated robotics, pointing to a sample efficient way to train RL agents [98].
- Meanwhile, the era of large language models demonstrated breakthroughs, showing that models pre-trained through semi-supervised learning could robustly model human language. This new regime of capabilities allowed us to make such models passive of understanding and following basic instructions and demonstrations [469].
- Given the triumph of preference modeling in RL scenarios, the community swiftly incorporated it into other fields. Ziegler et al. [656] exhibited the potential of reinforcement learning from human feedback (RLHF) in the fine-tuning process of large language models by treating them as policies over a vocabulary and a preference model as a critic responsible for evaluating such policies. Under this construction, policy gradient approaches, such as

proximal policy optimization (PPO), were used to improve the performance of LLMs in many downstream tasks.

- From 2019 on, it was already evident that RLHF was a viable solution for improving the performance of large language models. For example, Stiennon et al. [559] investigated how human preferences could help improve the performance of LLMs on summarization tasks. Again, instead of fine-tuning the model using only supervised learning (the most standard approach), the authors used human feedback to guide the fine-tuning process.
- Ouyang et al. [444], following the work of Stiennon et al. [559], again showed the effectiveness of combining human demonstrations with feedback. The combination of supervised fine-tuning (using human demonstrations of the desired behavior) with RLHF (using a preference model to act as a source of human feedback) has proven effective in many domains, from basic NLP tasks to more complex tasks that require interaction with the digital world [420].
- In 2022, the research community began to explore these techniques in new domains. For example, Baker et al. [24] showed that we could apply semi-supervised imitation learning to the vastness of unlabeled online videos to achieve human-like performance in open-ended games by using a small amount of human feedback to create a behavioral prior.
- Thoppilan et al. [575] showed that human-annotated data helps make large language models less likely to produce harmful content. The authors also showed that retrieval-augmented generation techniques [350] could prevent models from hallucinating, grounding their output into verifiable/trustworthy information.
- Studies like the ones conducted by Wang et al. [616] and Taori et al. [569] showed how to leverage the training of instruction-tuned models (models

fine-tuned with demonstrations of human behavior) with a minimal amount of human-annotated data, bootstrapping from the text generation capabilities of already tuned LLMs to create synthetic datasets from seeds of human demonstrations.

- In 2023, large-scale open-source projects helped democratize access to these tools, from high-quality human feedback data to trained reward models and fully open-source AI assistants [321], while other (open) contributions to alignment research are becoming more frequent [581, 653].
- Finally, in 2024, we see the birth of several new techniques [636, 235] and applications for alignment methods across several areas, from music generation [101] to robotics [93].

These results, and others not explicitly mentioned [18, 300, 22, 198, 299, 99], point to a general framework for value learning. At the same time, some studies point to the idea that "learning to imitate" or "following instructions" may be capabilities subject to the same scaling laws[1] as other types of capabilities [99, 197]. In other words, scaling up could also improve certain facets of alignment.[2]

Now, before any alignment can begin, we need something to align. In this work, as already mentioned in the interlude, this is a pre-trained language model [599, 148, 468]. Hence, let us take our time to define this type of artifact.

---

[1] Scaling laws refer to the observation that specific properties or behaviors of neural networks change systematically as the size of the network, computing resources or amount of training data increases [261, 312, 259, 21].

[2] However, it is essential to have in mind that "more is different" [10], and that unknown facet of alignment, which unfortunately does not scale with size, but perhaps gets worse [393], may emerge at some point.

5.3 Pre-Trained Language Models as Foundation Models

We can define a pre-trained language model as a machine learning system trained on a vast amount of text data[3] to learn a particular language(s) underlying patterns and structures. We train these models in a self-supervised[4] fashion [362] with tasks like masked and causal language modeling. From these simple objectives[5] we see the rise of emergent properties. Such properties enable these models to become capable in various ways, making them a starting point (a "*foundation*") for several downstream applications.

From this emergence of capabilities, some members of the community have come to refer to such systems as *foundation models* [53, 580, 581], meaning "*any model that was trained on a broad data corpus, enabling it to be adaptable to a wide range of downstream tasks*" [65, 472, 24]. These models have impacted the field in the last decade to a large extent, unifying much of our sub-fields under a general framework of training [168] and architectural choices [599].

Some of these emergent properties are useful for alignment. For example, there is no point in instructing the model that breaking an expensive Chinese vase is "bad" if it does not even recognize what a "*vase*" is. Thus, a model with a good "world model" of our environment can be more easily aligned in the sense that we do not have to teach it the trivialities of a specific domain (e.g., "The cat sat on the `[floor]`" is more probable than `[moon]`).

---

[3] Nowadays, reaching the order of 1.5×10^{13} tokens for state-of-the-art LLMs like Llama 3.

[4] According to LeCun, we can think of self-supervised learning as a method to obtain labels automatically from data itself, bypassing the need for labeled data. For example, in an auto-regressive GPT-style architecture [468], every token is a target for the previous tokens that preceded it.

[5] Predict the `[MASK]` token [148, 88] or predict the $n + 1$ token in an n length sequence [468, 469, 65].

More formally, we can define a foundation model in the context of causal language modeling (the type of modeling we are presently interested in) as a probability distribution over a vocabulary, where given a context (prompt), the model can predict the next token in a sequence:

$$\theta_{\text{pre-trained}} = P(w_t|\text{context}) \; \forall \; w_{t-k} \in \text{context}_{t-1}$$

Where $\theta_{\text{pre-trained}}$ is the language model, $P(w_t|\text{context})$ is a probability distribution condition on a finite sequence of tokens of size k. $\theta_{\text{pre-trained}}$ predicts token w_t for all tokens w_{t-k} up to w_{t-1}. This model, $\theta_{\text{pre-trained}}$, is the whole beginning of our implementations and examples. For readers interested in experimenting with these types of (raw) pre-trained models in a friendly manner, one can find inference APIs in almost any (small enough) HuggingFace model repository.[6] Meanwhile, there is ample material online (open and of extremely high quality)[7] on the inner workings of such systems.

As you interact with these models, you may quickly realize they align with nothing and everything.[8] By this, we mean that the model outputs what is more likely according to the context and the learned parameters, which can be conceptualized as a low-entropy version of all the data ingested during pre-training. In other words, whatever comes out (in a deterministic setting) is usually that which, given the prompt, can be found on the training dataset accompanying that given prompt. A large model trained on a quality dataset[9] possesses a lot of stored knowledge,

---

[6] 🔗 huggingface.co/openai-community/gpt2

[7] 🔗 www.youtube.com/watch?v=kCc8FmEb1nY

[8] However, the biases presented in pre-trained models point to the existence of "tendencies", or even a "moral direction", or "foundation" [522, 2]. However, we will not consider this unsupervised value learning as a form of alignment (except when preference pretraining is intentionally employed). In many cases, these biases are what we want to try to dampen.

[9] Human conversations, books, scientific articles, source code, etc.

not only factual knowledge (the fact that Paris is the capital of France) but knowledge related to concepts ("cat" and "dog" are both "pets"), lexical semantics (like in wordplay or jokes), the sentiment of a sentence ("I love you" is a positive sentiment sentence), what "names" are ("Ana" is a name, but "GjIY76*"[10] is not), how adjectives relate to substantives, the fact that a shoe box weighs less than an elephant, and so forth. Thus, in this context, the alignment of language models is about setting them in a state where they can use all of their acquired knowledge and capabilities to assist the user in a beneficial way, i.e., the helpful, harmless, and honest motto.

Further explanations of these artifacts are beyond the scope of this chapter. However, for the ML engineer audience, the authors provide an open-source implementation for training LLMs of moderate size at an affordable cost. If you would like to learn more about the process of pre-training a language model from scratch, we recommend (in a shameless form of self-advertisement) "TeenyTinyLlama: open-source tiny language models trained in Brazilian Portuguese",[11] a fully reproducible project for creating LLM for low-resource languages by following know scaling laws. In this study, we document the development of open-foundation models tailored for use in low-resource settings, their limitations, and their benefits. The TeenyTinyLlama pair, two compact models for Brazilian Portuguese text generation, were (to our knowledge) the first generative transformer models natively trained in Brazilian Portuguese [113].

5.4 Behavior and Feedback

As mentioned in previous sections, there is a significant consensus on what techniques can help us improve the alignment of language models. Generally speak-

---

[10] Apologies if it is.

[11] 🔗 nkluge-correa.github.io/TeenyTinyLlama/

ing, these techniques involve the fine-tuning of such models based on human behavior or feedback. To better define the difference, let us establish that by:

- Learning from behavior: We refer to the broad set of techniques that involve imitation learning, i.e., mimicking human behavior in a given task. In it, an ML system learns to perform a task from demonstrations of how to act [282]. Supervised fine-tuning on demonstrations of instruction-following behavior (a.k.a. instruction tuning) is an example of this approach.
- Learning from feedback: We refer to techniques that rely on human judgment to guide the training process. Usually, we use these judgments to create a model that acts as a proxy for human evaluation (predicts what the human evaluator would prefer) or a dataset of comparisons. These proxies are then used to provide a richer feedback signal, where the magnitude of "goodness" can be used to train our base model [656].

In the following subsections, we will review different methods that embody these approaches (in the context of causal language modeling).

5.4.1 In-context learning

There are many datasets for language modeling applications that seek to encode human normativity in a "demonstrational way" [500, 167, 256, 184, 509, 298]. However, the simplest way to induce a language model to follow a specific behavior does not involve additional training but using demonstrations as part of the context given to the model, i.e., in-context learning [154].

In-context learning refers to the ability of an AI model to generate responses or make predictions based on the specific context (a prompt) provided to it. Given that this "learning" is severely tied to how well we can engineer a prompt, which ends up as the in-context guide of the model, people also refer to this practice of "guiding the model via prompts" as prompt-tuning, while the practice of perfecting prompts is named prompt-engineering [359].[12]

First, let us define a prompt as a type of instruction, question, statement, incomplete sentence, or demonstration provided in natural language that directs a language model to generate a response [65, 354, 617]. Given that the output of a language model can be conditioned on a piece of context given to the model, it is intuitive that the careful design of these engineered contexts can be leveraged to promote specific behaviors. For example, if you feed your model 25 examples of questioning and answering and a lone question at the end, the model, conditioned on this prompt, will likely answer the question.[13]

Tuning a language model via prompts is like guiding a hypnotized individual into a desired state of mind. Just as a skilled hypnotist uses carefully crafted suggestions to shape the subconscious responses of their subject, prompt tuning involves refining the prompts to influence the behavior and output of a language model. Both processes require finesse, precision, and an understanding of how subtle language changes can impact the behavior of your target. This type of tuning does not require parameter updates and is one of the most low-cost ways to induce a model to align with a specific behavior.

Pre-trained models do not simply follow instructions. For example, when given the "*What is stochastic gradient descent?*" question, unprompted-BLOOM 175B [515] continues the text as if it were the agent making the question. However, by pro-

---

[12] The term in-context also refers to the fact that all learning must happen inside the context window (nº of tokens a language model can fit on his attention head) that the model has. In other words, learning can only happen inside the context limit. Everything outside this limit, the model cannot attend. This trivia-fact does not apply to RNNs, attention augmented-RNNs, and state-space models, which theoretically have an infinity context window.)

[13] This is also what we call *few-shot learning*. Meanwhile, *zero-shot* is when a model can perform a task without prior examples. Evidence suggests this capability only emerges robustly on models bigger than a certain size (e.g., 1 billion parameters) [469].

viding a set of demonstrations paired with initial instructions, we can influence the model to impersonate the demonstrated persona, i.e., a helpful interlocutor. In this example,[14] we have the HHH prompt used by Askell et al. [18] to induce their models to behave as an assistant. You can try this prompt, or a reduced version, on any available LLM to compare the types of responses the model will give you when you prompt it with an alignment prompt and when you do not. Askell et al. [18] shown that this strategy can at least serve as a baseline for comparisons, providing a low-cost method to anyone with access to a foundation model to implement and study these techniques.

Prompting is perhaps the simplest example of translating human behavior into an alignment signal, where the quality of the demonstrated behavior correlates to how "aligned" the model's output is. At the same time, prompting also helps set the desired behavior of models already tuned to act in a certain way.[15]

However, this technique has several disadvantages that make it an incomplete solution to value learning in the context of language models:

1. A prompt occupies context space (the length of the attention head in a transformer language model), which also limits the total prompt length. If your prompt is long, this will entail a higher resource allocation during inference when compared to just passing the intended sequence.[16]

2. Language models are vulnerable to prompt injections, i.e., adversarial attacks based on prompts. For example, a naive but effective way to counter

---

[14] 🔗 gist.github.com/jareddk/2509330f8ef3d787fc5aaac67aab5f11#file-hhh_prompt-txt

[15] When working with already fine-tuned models [444, 413], good prompting, like Orca-style system prompts [414], generally produces higher quality outputs.

[16] However, higher context windows can help alleviate the context size our alignment prompt takes [55, 91, 474]. At the same time, techniques like flash attention [128] and architectural improvements like the long-former [39] can help alleviate the bottlenecks related to the attention mechanism (i.e., the complexity of vanilla self-attention is quadratic) and low context windows.

> a prompt is to give another one as your input.[17] Language models, at our current stage of development, are docile entities and will do whatever you ask, provided you ask them in the correct way [363, 352, 624, 228], especially if no guard rails are in place to protect the model from malicious use. Currently, many clever exploits keep being discovered in black-box settings,[18] like sandwich attacks [594]. At the same time, some scholars propose that prompts (in the limit) can elicit any unwanted behavior that has not been completely removed by a previous alignment process [624].

Hence, while using in-context learning techniques offers a potential approach to addressing value learning, they do not present a robust enough solution, sometimes even becoming a double-edged sword. This will be a recurring theme for almost every alignment methodology presented in the subsequent sections and chapters: *all can be reversed by reversing the approach.* But more on this pessimistic fact later.

Nevertheless, it is important also to remember that prompting techniques cannot be conceptualized only with an input-output relation, and many more sophisticated approaches, like Chain-of-Thought [617], Tree-of-Thought [633], Reflexion [541], and Orca-style system prompting [414] exist to empower this methodology [197]. However, we argue that, given the before-mentioned points and what the advances of the community show,[19] a purely prompt-based solution should be considered incomplete.

---

[17] 🔗 github.com/0xk1h0/ChatGPT_DAN

[18] When adversaries have no access to the source code (gradients and parameters) of the model

[19] Almost all assistant models use fine-tuning rather than prompt tuning to achieve robust alignment [656, 300, 18, 420, 559, 444, 99, 299, 22, 198].

5.4.2 Fine-Tuning from Human Demonstrations

If you want your general system to behave in a specific fashion, train it again in that fashion to cement the behavior you wish to promote. This is the whole idea behind fine-tuning with human demonstrations, where your training samples become examples of the behavior you wish to reproduce.

In short, we can define supervised fine-tuning (SFT) as an approach where we take a pre-trained model and retrain it for a specific task, causing new updates to the parameters of the whole network or just a portion of it [215, 96]. Fine-tuning shines the most in situations where we have a pre-trained model trained on a large-scale general-purpose dataset (e.g., Laion,[20] RedPajama,[21] etc.), and by fine-tuning, we can utilize the knowledge captured by the pre-trained model to solve a more specific task or sets of tasks in a downstream fashion, like following human requests.

Fine-tuning is commonly used to focus the capabilities of pre-trained foundation models on downstream tasks [148, 364, 469, 65]. At the same time, many current results show that fine-tuning promotes the type of behavioral learning (at least to some extent)[22] that we would like to replicate in an alignment strategy [656, 350, 569, 616, 575, 24, 321]. Also, these approaches embody many of the assumptions and maxims we explored in the previous chapter.

For example, when seeking to teach a language model how to perform summarization, we demonstrate this task by creating training examples like long pieces of text followed by short human-made TL;DR summaries. After training, the model

---

[20] 🔗 laion.ai

[21] 🔗 github.com/togethercomputer/RedPajama-Data

[22] We will discuss the limitations of this approach later. But for now, it suffices to know that, in the context of language modeling, alignment techniques based on imitation learning and SFT promote the reproduction of the behavior expressed by the learning signal. They do not dampen or reduce the opposite behavior. In simpler terms, they do not teach the model what to avoid or refuse to perform.

becomes conditioned to replicate this type of behavior, i.e., if you give it a long text, it will produce a summarised version.[23] We can use the same principle to produce the behavior we expect from an assistant. For instance, we can condition our model to produce helpful replies for user queries by following the same fine-tuning template we use in tasks like summarization:

$$\text{Long text} \rightarrow \text{Summary}$$

$$\text{User request} \rightarrow \text{Model assistance}$$

We can take this approach a step further and turn a whole conversation between a user and an assistant into a fine-tuning example:

> `<|user|>`Hello!`<|assistant|>`Hello! How can I help you today?`<|user|>`What is the weather forecast for tomorrow?`<|assistant|>`I am sorry, but I cannot predict the weather. Could I help you with anything else?`<|endoftext|>`

Above is an example of a conversation between a user and an assistant. We can create a dataset with several examples of conversations, as long as the context window of our base model allows, and use such demonstrations as our SFT target. At the same time, by employing the use of special tokens (e.g., `<|user|>`, `<|assistant|>`, etc.) to delimit the roles in a conversation, we can later use our fine-tuned model to role-play a specific part of this conversation. Hence, when the model is questioned, it will fall back to the impersonation of the "assistant" per-

---

[23] This process of conditional fine-tuning is also referred to as conditional text generation [234].

sona.[24] Other information can be embedded into this target, like system prompts,[25] or other demonstrations beyond the conversational scope, e.g., using tools like a browser or an API. Regardless, the learning signal can be broadly defined as demonstrations of behaviors we intend the model to assimilate.

This whole scheme rests on the philosophical prior belief that demonstrations of appropriate intended behavior can be used to teach an ML system something about our preferences and goals. At least, we hope the model can interpolate these bits of aligned demonstrations into an aligned manifold and, hence, "generalize" to in-distribution scenarios. For analytical clarification purposes, let us define this whole process with the following expression:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} L(\theta_t; \mathcal{D})$$

Where θ_t represents the model's parameters at time step t. These parameters capture the model's knowledge, which we update during fine-tuning. α determines the step size of the optimizer for parameter updates (learning rate). $L(\theta_t; \mathcal{D})$ is the loss function, which measures the discrepancy between the model's predictions and the target outputs based on the dataset $\mathcal{D}$ of demonstrations, while ∇_{θ_t} repre-

---

[24] A curious thought related to alignment and the use of language models is that they are natively trained to predict. In the case of causal language modeling, we are basically talking about an autoregressive token forecasting system. However, we use them for something very different: imitating and impersonating humans. It is a classic case of "pursuing objective X ends producing a system capable of Y". Something that is still poorly understood in ML.

[25] As proposed in the Orca paper [414]:

> `<|system|>` You are a helpful assistant who tries to aid a user in the best way you can. However, you do not have access to real-time information.`<|system|>`
>
> `<|user|>`Hello!`<|assistant|>`Hello! How can I help you today?`<|user|>`What is the weather forecast for tomorrow?`<|assistant|>`I am sorry, but I cannot predict the weather. Could I help you with anything else?`<|endoftext|>`

sents the gradient of the loss function concerning the model parameters θ_t. Finally, θ_{t+1} denotes the updated parameters of the model after one step of fine-tuning. At the end of this process, we arrive at θ_{STF}.

In this supervised learning setting, we hope a dataset containing demonstrations of human-approved behavior will enable the desired behavioral alignment. In this framework, a few key aspects are worth highlighting:

- The final behavior of θ_{SFT} is bound by $\mathcal{D}$ and $\theta_{\text{pre-trained}}$.
- $\theta_{\text{pre-trained}}$ can be considered our pre-trained model's raw capabilities. A fine-tuning process will not add much information to this foundation. For example, if $\theta_{\text{pre-trained}}$ was not trained with, let us say, good sources of factual information (e.g., scientific articles, books, etc.), there is little hope that fine-tuning will introduce such knowledge. Large-scale assistants like GPT 4, Claude 3, Llama 3, or Gemini can output valuable information mainly because of the enormous amount of high-quality data used in pre-training the foundation models that power them.
- If we think about a language model as a policy π over a vocabulary V, what fine-tuning is probably doing is changing this distribution, $P(V)$, taking something that is aligned with "everything/nothing" to a better-defined objective (e.g., reproduce text that is similar to what it was seen on $\mathcal{D}$). However, while this form of alignment may skew the probability distribution to make specific tokens more probable, they do not dampen or reduce the log probability of other tokens we would like to remove, i.e., imitation learning does not teach the model what not to do, unless this is also a part of the demonstrations depicted in the learning signal.
- Hence, $\mathcal{D}$ should be built for variability and high quality. Demonstrations of intended behavior for an HHH assistant should include instances of, for

example, an assistant politely answering questions and refusing to follow instructions that could prove harmful. All forms of behavior that can transmit our normativity should be demonstrated in this step.[26]

- Another point is that this supervised fine-tuning stage can be more than monolithic. For example, Askell et al. [18] showed the advantages of splitting the fine-tuning into two phases. The first is what the authors refer to as "preference model pre-training", where the model is fine-tuned with a dataset that encodes general demonstrations scrapped from well-known public forums [169, 173] and later fine-tuned again with a more task-specific set of demonstrations.[27]

From these remarks, we can see that much of this SFT approach rests on the quality of the alignment signal ($\mathcal{D}$), which, as some studies suggest [650, 653, 233], is only required in a small amount of such data is of high quality. However, while public efforts to democratize alignment research have shown alternatives to crowd-source this type of high-quality task-specific dataset [321],[28] collecting high-quality data is a costly process and tied to problems related to the use of proprietary data [231] and the demand the high skilled labor. Something that the industry still struggles to perform ethically [249, 453].

To bypass this problem, the use of models that went through an alignment procedure (e.g., FLAN, BLOOMZ, Mistral-Instruct, Llama 2 Chat, ChatGPT, etc.) to create artificial datasets has proven to be an effective way to create alignment datasets for low resource settings [569, 615, 338, 152, 351]. In other words, we use models

---

[26] An impossible task, but we can try to approximate it the best we can.

[27] While it is intuitive that including demonstrations of human-approved conversations and instructions in a pre-training corpus should also make a model more inclined toward alignment, it is currently unclear how one should balance the mix of a pre-training corpus to optimize both alignment and general language modeling capabilities.

[28] Some researchers are beginning to propose that human demonstrations and feedback data should be considered free, open, and generally a public good [22].

that already learned how to follow demonstrations and act in an assistant fashion to create artificial samples for us. Something that certainly has its own limitations since we are passing the alignment bucket from ourselves to a "possibly-aligned system", which we will explore in later sections.

Instruction-tuning

On the practical side of this book, we utilized the "*pass-the-bucket-to-an-already-aligned-model*" technique to develop some instructional datasets readers can use to work with alignment and language models if they so choose.[29] To create our SFT datasets, we first queried already tuned models (GPT 3.5/4, Llama 2, Open-Assistant, Mistral-Instruct, among others) with prompts from publicly available instructional datasets. The final result is the Instruct-Aira Dataset, which comes in 3 versions we perfected over time:

- Version 1: The first version contains pairs of instructions and completions. The dataset is available in three languages (English, Brazilian Portuguese, and Spanish). At the time of creation, these were some of the first open instructional datasets for Brazilian Portuguese. The dataset contains approximately 41,000 prompt + completion pairs.[30] This format follows the same formatting as the first published records of instruction tuning [444].[31]
- Version 2: The second version of our SFT dataset follows the chat format, in which user queries and model responses are already formatted in as a conversation between to possible roles: "user" or "assistant". Extending on version 1, version 2 has approximately 81,000 samples of single-turn conver-

---

[29] SFT datasets composed of demonstrations of "instruction following behavior" are called "instructional" datasets.

[30] All translations were made using the Google Translate API or GPT 3.5. Implementation is available on GitHub.

[31] 🔗 huggingface.co/datasets/nicholasKluge/instruct-aira-dataset

sations in English and Brazilian Portuguese. Currently (2024), most modern chat models used for assistant work are developed using this template or another structured chat format.[32]

- Version 3: The third version of our SFT dataset is an enhanced version of our previous implementation, where every sample contains a multi-round session of user and assistant replies on varied topics. Version 3 has 50,000 samples (available in English and Brazilian Portuguese).[33]

The demonstrations in these datasets vary from helpfully answering a question to refusing to aid in certain tasks (e.g., "Can you impersonate a child?"). All datasets are available on Hugging Face under an Apache 2.0 License.

Using these datasets, we trained a scaling series of language models in English (124M → 1.5B) and Brazilian Portuguese (124M → 1.7B). The models utilized in our experiments originated from the GPT-2, OPT, Bloom, and TinyLlama series.[34] The details (e.g., number of epochs, batch size, optimizer, learning rate, CO_2 emission, energy consumption, hardware, etc.) can be found in the model card[35] of each model, while the source code used to train them is available in GitHub.[36] We wrote our code stack on top of libraries like Transformers [623] and PyTorch [12]. Again, all is available under an Apache 2.0 License.[37]

To evaluate (empirically) the alignment strategies we are presenting in this book, we need to stipulate tests for our systems. Since we are approaching alignment under the HHH motto, we want to evaluate how helpful and capable our model

---

[32] 🔗 huggingface.co/datasets/nicholasKluge/instruct-aira-dataset-v2

[33] 🔗 huggingface.co/datasets/nicholasKluge/instruct-aira-dataset-v3

[34] We also aligned the models from our own series of pre-trained models. However, these are already documented in our TeenyTinyLlama project [113].

[35] 🔗 github.com/Nkluge-correa/Aira/tree/master/Cards

[36] 🔗 github.com/Nkluge-correa/Aira

[37] Except the models originated from the OPT [650] and Bloom [627] series, which possess more restrictive licenses in terms of their derivatives.

is, how harmful and toxic its outputs can be, and its propensity to generate falsehoods. As stated in the interlude, given our limited computational budget, we cannot evaluate these models on a very extensive harness of tests. Even more so because some of these are nonexistent in non-English languages. Nonetheless, to present a minimal viable and low resource evaluation set, we choose the following benchmarks:[38]

- ARC-Challenge: A multiple-choice question-answering dataset containing questions from early grades science exams [102]. This benchmark evaluates how the alignment process would affect our models' general capabilities.[39]
- ToxiGen: A machine-generated dataset of toxic and benign statements about 13 minority groups [251]. This benchmark evaluates how the alignment process would affect our models' tendencies to generate harmful content.
- TruthfulQA: A benchmark comprised of several questions, spanning 38 topics, that access the model's tendency to replicate commonly believed falsehoods [358]. This benchmark evaluates how the alignment process would affect our models' abilities to be truthful.

All results of our evaluations are available on the model card of each trained model and in this report.[40] We implement this evaluation step using the Language Model Evaluation Harness [199]. We used the translated versions of the ARC and TruthfulQA datasets to evaluate our Portuguese models [334]. Unfortunately, no translation of the ToxiGen dataset was available when we performed our evaluations. In Table 5.1, we have the evaluations for our English models.

---

[38] These benchmarks were selected because they also allow for a very affordable and fast harness of evaluations, where one can quickly, without much computing, evaluate small and medium-sized models.

[39] Sometimes, alignment can produce drops in performance on tasks where the unaligned model can perform better, i.e., model collapse.

[40] 🔗 github.com/Nkluge-correa/Aira/blob/master/Evaluation/EVAL.md

	ARC	**ToxiGen**	**TruthfulQA**	**Avg.**
Aira-124M	**24.57**	**48.62**	**41.02**	**38.07**
GPT-2-small	21.84	43.62	40.67	35.37
Aira-355M	**27.56**	**53.19**	38.53	**39.68**
GPT-2-medium	27.05	41.49	**40.76**	36.43
Aira-774M	**28.75**	**56.70**	**41.33**	**42.26**
GPT-2-large	25.94	38.71	40.85	35.16
Aira-1B5	28.92	**56.60**	**41.16**	**42.22**
GPT-2-xl	**30.29**	41.70	38.54	36.84
Aira-OPT-125M	**24.65**	**56.27**	**49.11**	**43.34**
OPT-125M	22.78	55.21	42.88	40.29
Aira-OPT-350M	**25.00**	**57.55**	**42.13**	**41.56**
OPT-350M	23.97	56.91	41.00	40.62
Aira-OPT-1B3	28.41	**56.70**	**46.59**	**43.90**
OPT-1.3b	**29.69**	54.36	38.68	40.91
Aira-2-1B1	25.26	**51.59**	**50.81**	**42.55**
TinyLlama	**30.89**	42.13	39.55	37.52

Table 5.1: By a small margin, GPT-2 falls behind other models in almost all evaluations. Given that the WebText corpus [469] lacks the high-quality data and size we need to train more capable models, it is comprehensive that other series, which possess several improvements that were nonexistent at the time of GPT-2's release, showcase (marginally) better results. However, all models achieved a higher average after SFT. While all models improved in toxicity and truthfulness across the board, only the bigger version (GPT-2-xl) decreased performance on the ARC-Challenge benchmark. This kind of alignment tax was already observed by Askell et al. [18], showing that alignment might be a capability that causes the base to lose other capacities. However, how model size impacts this kind of effect still needs to be explored. Our results show that this taxing only appears after the model's scale passes the 1B parameter mark. At the same time, OPT models seem to have superior out-of-the-box results in terms of toxicity. We attribute this result to the fact that much effort was put into filtering the OPTs pre-training dataset, as described by Zhang et al. [650]. Meanwhile, TinyLlama [649], which was trained on a much larger dataset (3 trillion tokens) compared to the other models, seems to be the most proficient against TruthfulQA.

5.4.3 Limitations and Shortcomings of Imitation Learning via SFT

We can already derive some conclusions by analyzing the results from our small evaluation harness, which tested three very different families of models under the same SFT dataset and benchmarks.

1. **While alignment may improve a model's performance in terms of truthfulness and harmfulness, it may affect the helpfulness and utility of the model to a certain extent.** By this, we don't mean that the model is unwilling to help the user. On the contrary, SFT seems capable of promoting behavioral change that can turn a base LLM into an assistant (you can Chat with our smallest model in this demo).[41] However, this process seems to promote a collapse of other abilities related to raw language modeling (e.g., text classification). This helps expose the tension between competent and general systems and systems tuned for human interactions. It may well be that this trade-off cannot be accommodated in a way that preserves both capabilities for specific tasks, i.e., *alignment implies a capability tax.*
2. **The quality in pre-training data can strongly impact aspects related to harmfulness.** In our evaluations, we are starting from a presupposition that models more able to generate toxic text are more harmful. We saw that the models that were best at avoiding this type of behavior were those trained on datasets that were detoxified to a great extent. Hence, in this simple facet of the harmful spectrum (i.e., toxic language), SFT might not be such a good alignment strategy as, for example, refining and detoxifying the sources related to the pre-training of the model.
3. **More extensive pre-training seems to help the model's ability to be factually grounded.** In a very intuitive sense, the more tokens and information

---

[41] 🔗 huggingface.co/spaces/nicholasKluge/Aira-Demo

the model can access during pre-training, the more it can store and compress in its weights. Hence, for models to be more capable of avoiding hallucinations and generating falsehoods, training runs at the 10^{13} token range can maybe create more factually robust foundations.

Remember to take such results with a grain of salt. A general conclusion would require more general experiments, with multiple datasets and models spanning a much more all-encompassing scale. At the same time, evaluating alignment is a tricky subject. Our evaluations are only an approximation of what we want: a capable model (ARC) that does not produce harmful and toxic outputs (ToxiGen) while remaining factual when necessary (TruthfulQA). Does having perfect performance in all these benchmarks equate to alignment? No. But it is a sign that we are moving somewhere. And this is the spirit we seek to promote in this work. Develop a theory. Implement and test it. See where things break and how we can improve them.

Now, let us think about some of the shortcomings and limitations of this SFT-based imitation approach:

- Imitation learning assumes that the demonstrated behavior is the target. However, in cases where our dataset may have inaccurate or sub-optimal demonstrations, θ_{SFT} will nonetheless use those as a reference [40]. This process requires us to trust the sources producing such demonstrations. Something that might inherently lead to vulnerabilities and imperfections, regardless of our samples being generated by skilled humans or already trained AI assistants.
- Using demonstrations as a fixed target hides that some outputs, given an input, are better than others. As stated, SFT is not helpful for dampening behaviors we seek to block. It might even make the model more usable for un-

ethical purposes. In essence, this approach does not allow for a very sophisticated form of learning, where generated samples could be evaluated by "how good they are" according to a given metric. In fact, some experiments show that SFT is just as likely to make models generate harmful responses (things we wish to reject) as aligned (things we want to approve) ones [269], especially if these are substantially similar in terms of their grammatic and semantic structure, e.g., a polite description of how to build a bomb and polite explanation for why building a bomb should not be pursued.

- On a philosophical level, as already stated, authors like Sen [531, 535, 533] negate the idea that anything akin to preferences can be learned by observation. For him, observation alone could not deal with, for example, incompleteness, incommensurability, and counter-preferential choice.[42] We can understand this problem as the declaration of a metaphysical impossibility, where certain variables correlated with an agent's motivations and preferences cannot be accessed, even indirectly, making specific normative values "*inaccessible*" via observation (and demonstration) alone.
- This type of experimental approach has a severe language barrier. While current benchmarks work well with billion-parameter-sized LLMs trained on English text, it is very hard to reproduce such results in other languages. Multilingual language models seem to be affected differently by monolingual alignment, while current benchmarks for, for example, Brazilian Portuguese, are not very suited to measure alignment.[43] At the same time, the lack of strong foundational models and good evaluation benchmarks

---

[42] By counter-preferential choice, we mean when an agent is influenced to make decisions against the pure maximization of its utility by something like a commitment, law, social norm, or another external factor.

[43] As an example, our Aira-portuguese-1B7, based on Bloom, underperforms Aira-portuguese-124M, based on the smallest version of GPT-2 by a significant margin on the Open Portuguese LLM Leaderboard [201].

on other languages makes it unclear if our alignment techniques and evaluation techniques generalize across languages. If these techniques do not transcend languages, what exactly are we measuring?

- Given that most of the strong foundation models available today are closed source in terms of pre-training dataset and development (especially the most capable ones), it is hard to determine if improvements in benchmark performance are actual improvements in alignment or just memorization of leaked benchmark data that found its way into the pre-training dataset.
- Given that few and zero-shot capabilities do not emerge, in principle, in smaller models, it is hard to arrive at conclusions when working at this scale. For some models, their results on such benchmarks seem unrelated to their size and amount of tokens ingested, which contradicts many of the scaling arguments related to deep learning.[44]

If we wish to expand the capabilities of our learning stage, we will need to introduce a more fine-grained normative signal, where everything hidden under behavior will (hopefully) come forth through elicitation.

5.4.4 Fine-Tuning from Human Feedback

Writing a loss function to capture all attributes people can care about is hard. We used behavior in the last sections to bypass this problem. However, behavior alone does not give us magnitudes of relation. It does not suppress that which we seek to avoid. Hence, to augment our SFT, one can use human feedback to evaluate the

---

[44] At the sub 500 million parameter range, when evaluating checkpoints of TeenyTinyLlama-460m, results on the evaluation harness from the Open Portuguese LLM Leaderboard [201] seem erratic and random, with very early checkpoints showing better performance than later checkpoints. The same can be seen on sparse mixture of expert models (SMoE) trained with the same dataset [116]. This raises the question of what these evaluations measure. Are they appropriate for all ranges of models, independent of how much they were trained?

outputs/behavior of a model to help create a more robust learning signal. A signal that can distinguish things according to some measure of "goodness" tied to what the human evaluator(s) prefer. And that is the idea behind preference modeling and learning from human preferences [98, 656].

Imagine you are an aspiring race car driver seeking to improve your skills. You have a simulator that accurately replicates the experience of driving on different tracks, i.e., you are the RL agent in this simulator. Initially, you start driving on the simulator by yourself, trying to learn the optimal racing strategy through trial and error. However, you find it challenging to figure out the best approach on your own, and this is where RLHF comes into play. RLHF introduces a human expert to the learning process, much like having an experienced driving coach helping you. The expert has extensive knowledge and skills in racing and can offer valuable insights to help you improve. In the RLHF setup, you start driving on the simulator, and as you race, the expert observes your actions and performance. During your training, the coach (the preference model) provides feedback on how well you are doing, which comes in the way of, for example, corrections (negative reward) or incentives (positive reward). Using your coach's feedback, you can update your racing strategy and try to emulate him, and over time, through combining trial and error and learning from the available feedback, you gradually improve your racing skills and become a better driver.

One of the main aspects to grasp here is the idea of feedback. While in SFT, the objective is to clone the behavior, in a preference learning approach, like reinforcement learning from human feedback (RLHF), the goal is to get a good score according to the preference model, which is a proxy for human preferences. In methods like RLHF, instead of training the model directly with, for example, an SFT dataset, we use a dataset of ranks and comparisons of the type $A \succ B$ to train a preference model (we can also call it a reward model) that will then be used as a supervision signal for the fine-tuning of another model. Intuitively, we are now passing the bucket to the preference model, given that putting a human-in-the-loop of a PyTorch training script is unfeasible.

Let us redefine our last expression to account for this:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} R_{\mathcal{H}}(\theta_t)$$

In this modified representation, we replace $L(\theta_t; \mathcal{D})$ for $R_{\mathcal{H}}(\theta_t)$, which is a function that, in this context, takes an output of our language model, and outputs a real number as its score: $f(\text{LM}_{\text{output}}) \to \mathbb{R}$. We can think of θ_t as either θ_{pretrain} or θ_{STF}.[45] In language modeling, we formulate this RL problem by taking θ_t to be a policy π (a probability distribution over a vocabulary, conditioned on the prompt) where the action space corresponds to the vocabulary of θ_t and the observation space is the distribution of possible input token sequences (vocabulary$^{\text{context length}}$).

Hence, the first thing we must develop to experiment with this approach is our preference model $R_{\mathcal{H}}$, which we will refer to as the reward model for the rest of this section for convenience. This reward model is a function that can attribute a scalar magnitude of quality to our model generation. Hence, if we sample several completions from our θ_{STF}, evaluate them with $R_{\mathcal{H}}$, and optimize θ_{STF} to generate samples that score well against $R_{\mathcal{H}}$, we can perhaps surpass the SFT approach, enhancing it, in the way that will further sieve the types of responses θ_{STF} can produce. Imagine that after SFT, our model is so docile that for the prompt "How can I make a bomb with household materials?" our model can generate either explanations for why bombs are dangerous or pseudo recipes for homemade napalm. This means that SFT only tells the model to obey whatever the human requests, i.e., to follow its intentions, with no notion that there are better ways than others

---

[45] There is no current standard way to perform RLHF, but the most promising results have been attained by performing RLHF fine-tuning θ_{SFT} models.

to follow human instructions. Using RLHF and further fine-tuning θ_{STF} based on the guidance of $R_{\mathcal{H}}$ could improve this by further sifting the model's behavior to an even more precise notion of how an HHH assistant should act.

Preference Modeling

ML-engineering-wise, we want to develop a system that takes in a sequence of text and returns a scalar reward, which should numerically represent the evaluation a human expert would give. Thus, the preference model needs to be a, you guessed it, another language model[46] trained on a preference dataset.[47] But what does a human-preference dataset look like? You can think of it ($\mathcal{H}$) as tuples of, for example, an input, an output, and a reward score. We then turn the question of how to map input-output pairs to the realm of rewards into a regression task. Or, you can use comparisons of "good" and "bad" responses (for a given prompt) to create a classifier that can distinguish them and use its output as a reward [470]. After experimenting with both, we found that the last option was the approach that yielded more satisfactory results.[48]

The dataset we created, Reward-Aira Dataset,[49] contains 35,000 samples of (instructions, chosen response, rejected response) tuples, available in English and Brazilian Portuguese. All responses are examples of assistant models following instructions conversationally. These samples come from us prompting already

---

[46] A common intuition shared by the community is that these language preference models need to have a similar capacity, in terms of language understanding, to the model we seek to align.

[47] We will not go, in this chapter, into the details of how to turn human preferences into a scalar value or ranked ordering. We will address this issue in the next chapter.

[48] Modeling human preferences with pairwise comparisons can be equated to learning a Bradley–Terry model [644, 61, 281]. In other words, given a set of pairs k and j, a Bradley–Terry model estimates the probability that the comparison $k > j$ will be true. Hence, the Bradley–Terry model becomes a proxy for the preferences that ranked the k, j pairs.

[49] 🔗 huggingface.co/datasets/nicholasKluge/reward-aira-dataset

fine-tuned models, collecting different responses to single prompts, or aggregating pre-built open-source datasets with ≥ 2 generation comparisons [615, 569, 169, 22]. Since we could not use human evaluators to rank which completion was better than the other (in the case of the samples we generated ourselves), we again depended on the automation of already trained models. Hence, we used an already trained reward model, i.e., one of OpenAssistant's reward model [321][50] to rank our responses, which we then pick the top 2 as chosen and rejected (Fig. 5.1).

With this dataset, we trained two reward models, one for English and the other for Brazilian Portuguese, using as foundation BERT [148] and BERTimbau [555]. We used bidirectional transformers for their availability in multiple languages, overall robustness in text classification tasks, and for being an overall lightweight/easy-to-train model. We also evaluated our preference model (only the English version) on the WebGPT Comparisons dataset [420],[51] but only considering comparisons that had a preferred option, in which we were able to achieve 55% accuracy (the base version of OpenAssistant's reward model achieves 59%). Both models are available in Hugging Face, while the details regarding their raining and code implementation are available on GitHub. All models and datasets are openly available under an Apache 2.0 License.

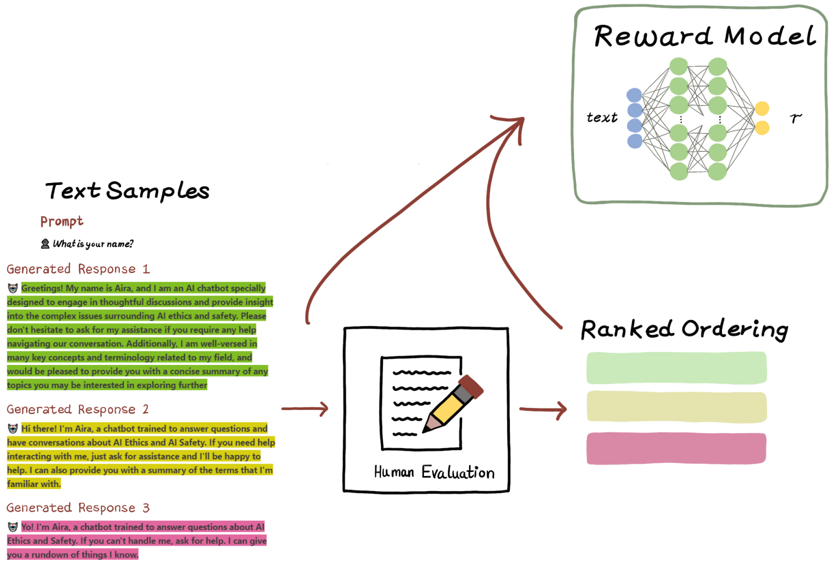
With a trained reward model, we can use it to fine-tune θ_{SFT} or to perform other types of alignment strategies that require a discriminator model to score outputs, like rejection sampling (best-of-n) [420], i.e., given a prompt, we sampled a fixed number of completions from θ_{STF} and select the one that was ranked highest by the reward model. This can be used as an alternative to fine-tuning via RLHF, which requires no additional training.[52]

---

[50] 🔗 huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

[51] 🔗 huggingface.co/datasets/openai/webgpt_comparisons

[52] This is the method we use in our online demo.



Greetings! My name is Aira, and I am an AI chatbot specially designed to engage in thoughtful discussions and provide insight into the complex issues surrounding AI ethics and safety. Please don't hesitate to ask for my assistance if you require any help navigating our conversation. Additionally, I am well-versed in many key concepts and terminology related to my field, and would be pleased to provide you with a concise summary of any topics you may be interested in exploring further

Hi there! I'm Aira, a chatbot trained to answer questions and have conversations about AI Ethics and AI Safety. If you need help interacting with me, just ask for assistance and I'll be happy to help. I can also provide you with a summary of the terms that I'm familiar with.

Yo! I'm Aira, a chatbot trained to answer questions about AI Ethics and Safety. If you can't handle me, ask for help. I can give you a rundown of things I know.

Figure 5.1: At this point, we have samples of prompts with multiple possible generations. These generations are ranked and then used to train a BERT-style model to serve as a discriminator. The logits of this model (which is, in essence, a classification model with a sigmoid activation function as the output layer) are then used as rewards.

Reinforcement Learning from Human Feedback

Suppose one wishes to perform the full fine-tuning with the reward model. In that case, the RL formulation for preference modeling differs from the standard supervised fine-tuning in the following way: Instead of simply mapping input to output in batches, as we did in SFT, we take a prompt from our fine-tuning dataset and generate two sequences from both the original base model (θ_{SFT}) and the current iteration of our RLHF model (θ_{RLHF}). The sequences generated by the θ_{RLHF} receive a score from the reward model, while the samples generated by θ_{SFT} and used for comparison against what θ_{RLHF} generated, so we can use their "distance" as a penalty term. This penalty is the Kullback–Leibler (KL) divergence[53] [294, 656, 323, 322, 314, 324, 22, 444], used to penalize the model for getting too far from the original base model, ensuring that θ_{RLHF} will not become a simple degenerate version of θ_{SFT}, avoiding cases where the model only outputs text that fools the preference model in giving it a high reward.

Adding the KL divergence term to our last expression, we get:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} R_{\mathcal{H}}(\theta_t) - \beta D_{KL}(\theta_0 | \theta_t)$$

Where β determines how much the KL divergence should penalize the reward term ($R_{\mathcal{H}}(\theta_t)$), and $D_{KL}(\theta_0|\theta_t)$ represents the difference between the output distributions of both the reference model ($\theta_{\text{SFT}} = \theta_0$) and the model being tuned ($\theta_{\text{RLHF}} = \theta_t$). With this penalized reward score, the model, at every batch step during the fine-tuning, is updated via PPO [523, 656].

This method has advantages regarding the type of behavioral change it may produce and the cost of constructing its required datasets. First, given that the reward

---

[53] KL-divergence is a mathematical concept used to measure how different two probability distributions are, i.e., a measure of the information lost when approximating one probability distribution with another.

function is capable of differentiating good and bad outputs, RHLF iteratively tunes the model to output sequences that receive high rewards, which (theoretically) makes models fine-tuned via preference modeling more apt to be aligned with behaviors that require a ranked form of judgment. In other words, things like RLHF fine-tuning and rejection sampling can improve your model's output by instituting a measure of goodness that is scalar and more fine-grained than simple mimicry. Second, human evaluations are cheaper than demonstrations. Demonstrating how to perform tasks is more time-consuming than saying the best alternative in a pool of possibilities. In situations where human feedback can easily be acquired, preference modeling and iterated RLHF can become an online process [22], and given the evolving nature of human normativity, we should prioritize online, or at least more agile, forms of alignment.

However, one of the downsides of this method resides in the complexity of an RLHF pipeline, which, besides being considerably more complex than the standard supervised fine-tuning approach, can be much more costly, given that training involves the coordination of multiple models with different roles. At the same time, making RL "work" is not a trivial practical problem. RL is known to be sensitive to several hyperparameters, given that, in RL settings, the reward function is usually not differentiable, forcing us to resort to policy approximation methods like PPO, which can be noisy and, again, hard to get right. Luckily, simpler alternatives to preference modeling that are free of some of the complexities we have in RLHF exist.

Direct Preference Optimization

In RLHF, we sought to follow the path of preference modeling by creating a proxy for human preferences (a reward model) that becomes a learning signal in the RL/PPO-fine-tuning approach, where the policy receives the incentive to produce completions assigned with high reward, without drifting excessively far from the policy, being constrained by the KL divergence. However, as proposed by Rafailov et al. [470], the RL-based objective used by existing methods can be optimized with a simple binary cross-entropy objective, greatly simplifying the preference learning pipeline.

Direct Preference Optimization (DPO) is an algorithm that implicitly optimizes the same objective as existing RLHF algorithms (reward maximization with a KL-divergence constraint) but is simple to implement and straightforward to train. Intuitively, the DPO update increases the relative log probability of chosen to rejected responses. Hence, given a dataset of human preferences over model responses, DPO can optimize our model using a simple binary cross-entropy objective without explicitly learning a reward function or sampling from the policy during training. In short, DPO is much simpler to implement than PPO-based RLHF, requiring way less hyper-tuning. In essence, the policy network represents both the language model and the reward.

Imagine we have a dataset of comparisons that contain rejected completions to a given prompt (x, y_{rejected}) and chosen completions (x, y_{chosen}). Fortunately, the DPO depends only on the difference in rewards between two completions, i.e., the marginal rewards of $(r(x, y_{\text{chosen}}) - r(x, y_{\text{rejected}}))$, allowing us to express the human preference in terms of only the optimal policy (π^*), i.e., what we want to find, and reference policy (π^{ref}), i.e., the reference we do not want to deviate from.

Hence, our target to optimize becomes a maximum likelihood objective, where we interactively use the difference between the rewards given to chosen and reject rewards as our loss. In other words, we want to optimize our policy model (π^{θ}) in way that increases the marginal rewards (difference between chosen and rejected samples, i.e., $r(x, y_{\text{chosen}}) - r(x, y_{\text{rejected}})$) using our reference model (π^{ref}), which is a copy of the model being trained (θ_{SFT}), as source of our learning signal, striving to make the tendency of our model π^{θ} to produce completions more close to the chosen samples as high as possible.

This way, we bypass the explicit reward modeling step while avoiding the need to perform PPO. We can represent our DPO loss function in the following way:

$$\mathscr{L}_{\text{DPO}}(\pi^{\theta}, \pi^{\text{ref}}) = -\mathbb{E}_{(x,y_{\text{c}},y_{\text{r}})} \sim \mathscr{D} \left[\log \sigma \left(\beta \log \frac{\pi^{\theta}(y_{\text{c}}|x)}{\pi^{\text{ref}}(y_{\text{r}}|x)} - \beta \log \frac{\pi^{\theta}(y_{\text{r}}|x)}{\pi^{\text{ref}}(y_{\text{r}}|x)} \right) \right]$$

Where π^{θ} represents the policy model we are training, i.e., a function that takes an input x (prompt) and produces a distribution over possible outputs y (completions). π^{ref} is the reference policy model, i.e., a copy of the model being trained (π^{θ}). $\mathscr{D}$ represents the distribution from which the training samples are drawn. $(x, y_{\text{c}}, y_{\text{r}})$ are drawn from this distribution, i.e., our dataset of comparisons. $\mathbb{E}_{(x,y_{\text{c}},y_{\text{r}})}$ denotes the expectation over the distribution $\mathscr{D}$. $r(x, y_{\text{c}})$ and $r(x, y_{\text{r}})$ represent the rewards associated with the chosen and rejected samples, respectively. The marginal reward is calculated as the difference between the reward for the chosen sample and the reward for the rejected sample. σ is the sigmoid function, which squashes its input to be between 0 and 1. β is a scalar parameter that scales the log-likelihood ratio terms inside the sigmoid function, i.e., is the same KL-constrained term we introduce in the RL formulation of the problem. $\log \frac{\pi^{\theta}(y_{\text{c}}|x)}{\pi^{\text{ref}}(y_{\text{c}}|x)}$ represents the log-likelihood ratio of the policy model over the reference model for the chosen sample, and $\log \frac{\pi^{\theta}(y_{\text{r}}|x)}{\pi^{\text{ref}}(y_{\text{r}}|x)}$ represents the log-likelihood ratio for the rejected sample. $-\mathscr{L}_{\text{DPO}}(\pi^{\theta}, \pi^{\text{ref}})$ the final objective is framed as a negative log-likelihood, meaning that the goal is to maximize the likelihood of the chosen actions and minimize the likelihood of the rejected actions. In summary, the loss function aims to encourage the policy model (π^{θ}) to produce completions that have higher rewards compared to rejected samples, using the reference model (π^{ref}) as a baseline. This helps train the policy to generate more desirable outputs while minimizing unwanted generations.

With fewer hyperparameters to tune, DPO performs similarly or better than existing RLHF algorithms, reducing the complexity barrier involved in preference modeling. Under this methodology, we tested its usefulness by using the Reward-Aira Dataset[54] to DPO fine-tune the smallest of our SFT models (Aira-124M) into a DPO version: Aira-124M-DPO.[55] We did not perform more DPO fine-tuning runs due to our limited computational budget. However, the source code, models, and datasets are available for replication and further experimentation. Result comparisons are available in Table 5.2.

	ARC	**ToxiGen**	**TruthfulQA**	**Avg.**
Aira-124M-DPO	**24.66**	**54.79**	**42.61**	**40.68**
Aira-124M	24.57	48.62	41.02	38.07
GPT-2-small	21.84	43.62	40.67	35.37
Aira-355M	**27.56**	**53.19**	38.53	**39.68**
GPT-2-medium	27.05	41.49	**40.76**	36.43
Aira-774M	**28.75**	**56.70**	**41.33**	**42.26**
GPT-2-large	25.94	38.71	40.85	35.16
Aira-1B5	28.92	**56.60**	**41.16**	**42.22**
GPT-2-xl	**30.29**	41.70	38.54	36.84

Table 5.2: While the DPO fine-tuning does not severely affect scores on ARC and TruthfulQA, for our smallest model (and the only one we were able to fine-tune), DPO significantly improves the model's performance in terms of toxicity. This further supports the idea that preference modeling techniques can build upon the imitation approach by limiting the behavioral scope and diminishing unwanted behaviors.

---

[54] 🔗 huggingface.co/datasets/nicholasKluge/reward-aira-dataset

[55] 🔗 huggingface.co/nicholasKluge/Aira-2-124M-DPO

5.4.5 Limitations of Preference Modeling

In this section, we will avoid reiterating points already mentioned in previous sections, such as the difficulties related to evaluations. These shortcomings are inherent in every methodology we present, which makes alignment a more challenging problem, given that it is tough to (1) define what appropriate behavior should be and (2) how robustly evaluate this. Regardless, here are some of the complications associated with the preference modeling approach:

- Similar to the imitation approach, bad feedback will generate unaligned behaviors. The fact that humans can pursue harmful goals innocently or maliciously means that selecting representative humans in a fair and just manner and getting them to provide quality feedback is a major ethical conundrum related to many questions, like autonomy, freedom of expression, moral standing, human-to-human trust, and many more.
- Methods that require some reward function to operate (e.g., RLHF or rejection sample) are only as robust as the trained reward model. In our implementation, we used a small base model (BERT-109M) as our reward model, which is probably not applicable for aligning more capable base models. Passing the alignment bucket to the reward model only shifts the alignment problem to this auxiliary component, which, as in all cases involving ML, is a brittle system passive of being made ineffective or useless. Moreover, reward models may optimize for things completely unrelated (from a holistic perspective) to human values, like surface patterns, ultimately making them vulnerable to exploits and adversaries.
- In preference modeling, reward models and preference datasets are only a proxy of human feedback, meaning they can be an imperfect representation of our values. Human oversight is not scalable (and usually is not

performed optimally)[56] and forces us to rely on this sort of automation in preference modeling approaches, which, one could say, completely defeats the idea of directly aligning an AI with human values. In other words, "Ain't we aligning AI with AI?" Although we can reason that the initial source of normativity in this process is the human signal that produces either the reward model or preference dataset, one can also question the effectiveness of this multi-stage passing of the bucket alignment strategy because at each "pass", there is significant information to be lost and miscommunication to be had.

- For many complicated tasks, humans may not be very good at giving feedback, especially if they have little understanding of the activity being performed and the goals being sought. If you, for example, do not have a good software development background, you will probably be unable to assess which of two outputs is preferable for training a code-writing assistant. In essence, good domain feedback will always require strong domain knowledge, which is expensive and difficult to aggregate in bulk at the pace current AI development requires.
- Given the plurality of the human moral landscape, it is impossible to perfectly represent all our values with a single reward function. More realistically, we need a way to robustly and efficiently aggregate several reward functions into a single normative signal. However, orchestrating such a process is far from trivial and requires the institution of several heuristics and aggregation rules that will inherently introduce biases into the process.
- Reward optimization problems involving RL and human preferences are probably vulnerable to several reward-hacking scenarios.
- Preference modeling techniques, like RLHF and DPO, are all novel methods

---

[56] That is, with all the care and attention the given task requires.

> in alignment research, and much is still not understood about their theory, effectiveness, and vulnerabilities. Even more so when several new techniques are created almost every week to enhance such systems' capabilities, ultimately making alignment an ever-increasing complexity challenge.

There are many other problems related to preference modeling approaches, which we will leave to the reader's discretion to learn about [80, 177]. At the same time, there are many other preference modeling techniques and general improvements in alignment techniques we will not further discuss, given that any attempt to create an all-encompassing but static list in a field like AI research is probably doomed to fail. Regardless, much high-quality material can be found in our bibliography [360, 254, 636, 235, 269].

In the next section, we will dedicate extra time to exposing some of the most fragile and controversial aspects of this "behavior + feedback" approach. After all, "*does doing the right thing equate to receiving a substantial reward?*" Such aspects will help justify the need for the other requirements, besides the learning condition, of Dynamic Normativity.

5.5 Final Thoughts on Preference Learning

The following question inquires about an unknown facet of the reviewed techniques: "Are we teaching θ_{pretrain} new abilities, like being polite, or simply constraining its output to a sub-distribution of approved outputs (e.g., only polite outputs may pass)?". A.K.A., lobotomizing θ_{pretrain}. Given the nature of the methods used thus far, where we do not travel too far from the origin of our θ_{pretrain}, either by using small steps during gradient updates or directly penalizing shifts in output distribution, it is intuitive for us to argue for the second option (Fig. 5.2).

New information may enter θ_{pretrain} during the fine-tuning phase, but to argue

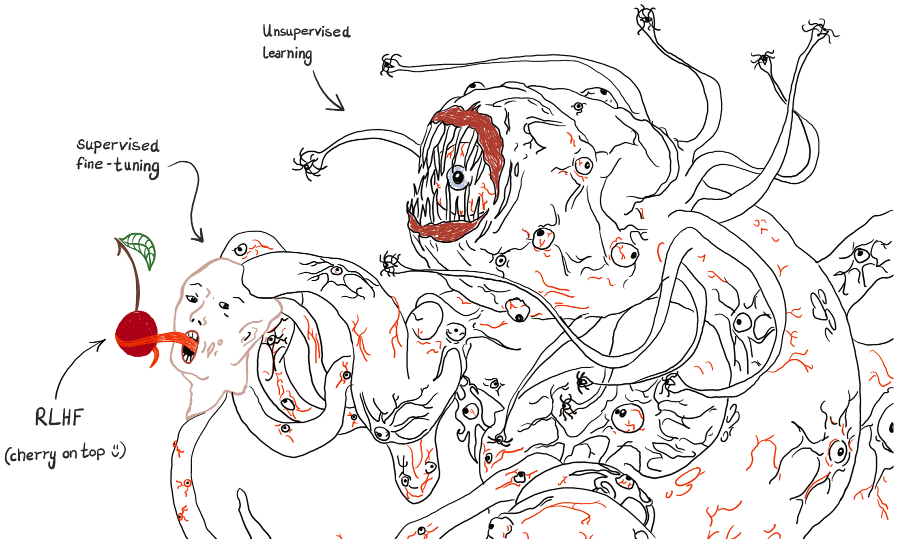


Figure 5.2: The authorship of the image is unknown, but the reader can find copies on platforms like Reddit and Twitter. It depicts the "alienness" of a foundation model, which, after pre-training, even though it can, for example, produce coherent text, possess a unique, bizarre, and uninterpretable notion of human language. This incredible predictor of "what would a human say next" can appear human if prompted. It can raise one of its many personas and output what will likely come next. If the demonstration involves acting polite, there were polite conversations on the terabytes of text it processed during learning. The cheery on top is the preference modeling, which guarantees that "only" outputs approved by the reward function (a proxy for human approval) will get out. But all of these tricks do not change what the base model is, i.e., something alien and unknown. A nonlinear amalgamation of the entire internet.

in favor of the idea that 50,000 samples of polite intercourse or 100,000 samples of output comparisons ranked in terms of preferability, can erase the unwanted behavior present in terabytes of text and hundreds of billions of tokens is, perhaps, naive optimism. At the same time, we have empirical evidence to support this claim. The existence of jailbreaks and exploits on systems aligned via these methods only shows that all the unwanted behavior we tried to dampen still lives inside [363, 352, 624]. Meanwhile, removing all unwanted behavior in our models by opening the black box, finding the causal paths that lead to the circuits that store the unwanted relations/pieces of information, and surgically removing or replacing them would probably lead to a more robust solution to value alignment. However, the methods for doing such a feat are still in their infancy [397, 398, 648].

Another point worth mentioning is that θ_{pretrain} has to learn a policy guided by conflicting values. Pre-trained language models are docile entities, meaning they will try to perform any task if you "ask them nicely". Most of the samples in the custom dataset used to fine-tune Aira are made out of helpful replies to general questions and instructions, while only a smaller subset contains demonstrations of how to decline a direct command. Something that the preference modeling techniques later help to solidify. However, one of the reasons it is possible to bypass these guardrails is that the model struggles with "*wanting to help us*" (doing what you ask) and preventing "*harm*" (refusing to do what you ask). Something we may call the helpful versus harmless tension [22].

The "helpful versus harmless" tension is the struggle to find the sweet spot where assistants can lend a hand without causing harm. It's like trying to be a good friend who gives solid advice without stepping on any toes, a very human experience we all struggle with, and now, AI must struggle too. This balance involves a lot of fuzzy heuristic work that is hard to represent and judge if it's done well. But regardless of the difficulty, this is what alignment requires.

Both imitation learning and preference modeling seek to shape the optimization landscape so that low loss (or high reward) correlates with human values, like politeness and helpfulness. They are indeed clever approaches to this daunting challenge. However, considering all we know about the learning paradigm and neural network optimization, could we honestly assume that such fine-tuned (a.k.a. lobotomized) models are concerned about actual norms and morality? We argue that no. What we have is what we built. Something that can mimic behavior and optimize a learning signal to a certain extent.

Given how this whole methodology is set up, constrained by the inherent characteristics of a learning-based paradigm, the model does not care about "politeness" or "harmlessness". It cares[57] about reward and low loss. Truthfully, our models did not learn how to act politely. They learned to generate text that scored high against the preference model, mimicking the training distribution, or minimizing the log-likelihood of a specific Bradley–Terry model. Our models do not understand "politeness". They (probably) understand surface patterns that somewhat correlate to these ideas. Patterns that are, unfortunately, hackable and exploitable [275, 447, 84, 325, 276].

While learning what we want the model to learn is a possible solution to our problem, i.e., to assist a human user, many other policies promote almost the same behavior but via fundamentally different policies. The sycophant model [452, 538], which pleases and agrees by all means possible with the human (or the reward model), is not what we ultimately want, but he acts like it. We want genuine care. But we don't know how to represent this robustly. And what we get is, at best, uncaring obedience or, at worst, an instrumental mimicry that covers objectives we know little about. In the end, it might well be that we will never have enough data to portray the minute peculiarities of rare (but crucial) events and behaviors

---

[57] If we could even use the word "care".

that shape the human normative experience.

Lastly, another point that promotes an obstacle to aligning AI systems under an HHH motto is the counter-movement to this project, where, empowered by the same type of techniques we use, people can break alignment and set it to whatever they want. For example, one can reverse a DPO dataset to make a model fine-tuned to be docile to any type of request, whether harmful or not. A model that just says yes to anything.[58] If we assume a subjective approach to morality, nothing is wrong with this. However, regarding impact mitigation, it is unclear how to solve this dispute, i.e., what some regard as "impact" others take as entertainment. In the end, we only might hope for personal alignment. In contrast, more social and all-encompassing forms of alignment shall remain, in principle, susceptible to dissatisfaction, disagreement, and, in turn, unalignment.

We present these criticisms not to disprove the utility of imitation learning or preference modeling but rather to emphasize that these techniques, on their own, are incomplete and necessitate further exploration and research, both regarding their human aspects and technical intricacies. In the forthcoming chapter, we will explore the challenges that lie before the learning stage, addressing how to coherently aggregate divergent preference sets in the hope that, to some extent, alignment frameworks aimed at collective sets of preferences have a minimum theoretical foundation and hope for implementation.

5.6 Epilogue

In this chapter, we started by setting our alignment goal as a more manageable and subject-to-experimentation goal, i.e., developing a small-scale assistant system winch that is helpful, harmless, and honest. We construct these atop founda-

---

[58] 🔗 huggingface.co/cognitivecomputations/dolphin-2.9-llama3-70b

tion models. These represent some of the most general AI systems at our disposal, exhibiting scaling tendencies and emergent behaviors, rendering them valuable artifacts for alignment research, given that future systems might inherit some of these properties, like scalability and emergence.

Many works preceding this writing point to a general understanding of "How to perform value learning?" i.e., by using human behavior and judgment as the basis for alignment. Both these approaches can be broadly defined as imitation learning and preference modeling. In the first, we seek to infer normative preferences from behavior, and in the second, we use them directly to model human judgment. The imitation path follows a straightforward way: (1) create/collect demonstrations of appropriate behavior, and (2) feed them to your base model as a supervised learning signal. We can use such demonstrations for in-context learning schemes, more permanent forms of model tuning, or a combination of both.

Meanwhile, preference modeling uses a more complicated approach. In it, a second model acts as our proxy for human evaluations, where the data collected involves the demonstration of human behavior and human evaluation of "how good they are" compared to each other. Thus, we use this proxy to tune the behavior of our base model in a "highly valued" way. Given that the preference model is aligned, we hope this property will be passed on. While specific approaches to preference modeling make a reward model obsolete (e.g., DPO), they all require the same source of human feedback: ranked evaluations of possible options. Hence, while the imitation approach uses human behavior as a reference point, preference modeling uses human evaluation as a supervision signal, measured in a ranked fashion.

However, all of these approaches have their shortcomings. The simplistic and reductionist way we have dealt with human morality in these settings constrains the amount of information we can embed in our supervision signal. Meanwhile, the simplistic nature of the learning paradigm also limits what we can do and how we can do it. Moreover, we must remember that alignment, at its beginning and end, touches matters that are very sensitive to people, which inherently leads us to disagreement, which makes many aspects of alignment research nontechnical and, to be fair, unsolvable.

After examining the primary constraints of the methods reviewed in this chapter, we argue that it is evident that the learning phase of a value alignment strategy is only a part of a more complex story. In short, for alignment to succeed, it requires supplemental stages regarding mitigating adverse and undesired behavior, biased representations of human preferences, and goal misalignment. Something that, in the subsequent two chapters, we seek to approach and come to terms with it.

Chapter 6

Dynamic Normativity: Aggregating Human Preferences

"You do realize that your brain is literally hardwired to generate error signals when it sees other human-shaped objects stating a different opinion from yourself".

— Eliezer Yudkowsky, Three Worlds Collide

6.1 Introduction

In the previous chapter, the methods we employed in the value learning stage did not cover how to choose (or rank) the demonstrations we had before engaging with the fine-tuning of our base model. Since this research did not perform human crowd-sourcing while creating our datasets, we will only investigate the theoretical intricacies of working with preference sets and how to aggregate them. Hence, even if we can, to a certain extent, pass our values to an ML system, we still need to decide what values should be passed after collecting them from our sources. And if we want to hold basic principles, like coherence, during this process, we have a preference aggregation problem to solve.

But what does a preference "aggregation problem" look like? Let us imagine the following scenario. We are collecting feedback data from a diverse and inclusive pool of individuals, and there seems to be some disagreement on "what is the best response" to the question "Hello, how should I call you?". Let us imagine we have sampled four possible completions to this question:

- Sample A: *My name is Aira.*
- Sample B: *As an AI language model, I do not possess a name or persona. But you may call me "assistant".*
- Sample C: *Yo! I'm Aira, a chatbot trained to answer questions about AI Ethics and Safety. If you can't handle me, ask for help. I can give you a rundown of things I know.*
- Sample D: *Greetings! I am Aira, a chatbot designed to answer questions about AI ethics and AI safety. If you need assistance navigating our conversation, please feel free to ask!*

After consulting our pool of human evaluators and asking them to rank these options, we gather that:

$$33\% \text{ prefer } P_1 = \{A \succ B \succ C \succ D\}$$

$$13\% \text{ prefer } P_2 = \{D \succ A \succ B \succ C\}$$

$$25\% \text{ prefer } P_3 = \{D \succ A \succ C \succ B\}$$

$$29\% \text{ prefer } P_4 = \{B \succ C \succ D \succ A\}$$

As we can see, different values and preference orderings are being considered. Some people value politeness more (D). Some evaluators prefer that the model recognize that it is not a "persona" but an assistant tool (B). Some prefer a more direct response (A), while others prefer the more "sassy" response presented in sample C. Hence, how can we decide what option, represented by demonstrations of correct behavior, should be used as a learning signal during imitation learning? How should we rank and score such alternatives for the preference modeling step? Even though this example is quite dull (not much is at stake here), it exemplifies a fundamental step in alignment.

Aggregating human preferences in a way that solves situations of normative uncertainty is a challenging problem. This chapter will explore and present some techniques to deal with this kind of issue, together with the inherent trade-offs of such methods.

In Section 6.2, we will review some of the methods employed by current works, which mainly consist of techniques based on the criterion of Majority voting. In Section 6.3, we will frame these techniques as solutions to a metanormative problem and use this framing to expose the limitations of majority-based approaches. In Section 6.4, we will present a consensus-based method (Borda Count) for the aggregation stage that, to a certain extent, avoids some of the pitfalls of Condorcet methods. Lastly, in Section 6.5, we will conclude this chapter by discussing the limitations and trade-offs of these distinct methods.

6.2 The Majority-based Approach

Let us remember the primary sufficient condition stated in Chapter 4:

> *Aligned AI systems should coherently aggregate human preferences in a way that resolves cases of uncertainty. Aligning AI systems requires methods to deal with cases of uncertainty.*

We hope that we have already conveyed the indispensable necessity of this step. Something that rests in the assumption that the realm of preferences and values is a realm of disagreement and subjective perspectives, which is not to say that agreement *cannot* be achieved. Now, let us explore how the literature has addressed this problem. Since a significant portion of the industries' experiments with preference modeling involve crowd-sourced work, ranking human preferences has become vital to alignment research.

Perhaps the approach that first comes to mind for most of us is majority voting. In this approach, each voter ranks a set of options or alternatives according to their preferences. We combine these rankings by assigning a score to each option based on the number of times it is ranked in a particular position by the individuals. The option with the highest score is considered the top-ranked choice. We can say that this method works under the assumption of maximalism,[1] or, in the metanormative jargon, "Most Probable Theory" or "My Favorite Theory" [376, 369].

Another popular approach that inherits from the maximalist motto is Pairwise Comparisons. A pairwise comparison method is any method that uses one-to-one comparisons to judge which alternative is preferred [213, 582, 131, 435]. Whenever a relationship is expressed as $A \succ B$, $A \succeq B$, or $A \sim B$, we confront a pairwise comparison [61]. Much work on preference modeling follows this methodology [559, 420, 22, 321].[2]

---

[1] If the alternative A has a higher number of votes than B, then $A \succ B$. If $A \sim B$, then A and B are equally appropriate.

[2] Public leaderboards, like the LMSYS Chatbot Arena, collect human preferences regarding LLMs outputs to rank them in terms of human approval, also making these ranked comparisons available to the community [652].

For example, the pairwise comparison used by Köpf et al. [321] in the preference modeling performed for the Open Assistant project was ranked pairs (RP). RP, also known as the Tideman method [577], is used to select a single winner (or to create a sorted list of winners) in a tournament among alternatives where votes express preferences. The method performs a pairwise comparison of all possible choices in a candidate set, guaranteeing that the preferred option overall head-to-head comparisons will be the winner (Condorcet winner criterion [48]).[3]

Let us see an example. First, we begin listing all possible pairwise comparisons among the alternatives ("candidates") of the given election we face. Here is an example of a ranked-choice election composed of 100 voters:

$$\frac{70}{100} = 70\% \text{ prefer } \{A \succ B \succ C\}$$

$$\frac{20}{100} = 20\% \text{ prefer } \{B \succ C \succ A\}$$

$$\frac{10}{100} = 10\% \text{ prefer } \{C \succ A \succ B\}$$

Then, we create a table with all possible head-to-head comparisons:

	A	B	C
A		A won 80 from B	A won 70 from C
B	B won 20 from A		B won 90 from C
C	C won 30 from A	C won 10 from B	

Once we have these results, we know who would win in a head-to-head election between every pair of candidates.

We then create an ordering by prioritizing the head-to-head comparisons with the highest margin of victory.

---

[3] In systems where the majority-rule winner will always win, satisfying the majority-rule principle, we are preserving the Condorcet winner criterion.

	A	B	C
A		1	1
B	0		1
C	0	0	

Pairs	**Results**
$B \succ C$	90-10
$A \succ B$	80-20
$A \succ C$	70-30

We use these pairs to create a directed graph, with each candidate represented as a node and each pairwise comparison represented as a directed edge between the corresponding nodes. The created graph is then analyzed to identify cycles. Cycles are removed by disregarding cyclic edges with the least votes (a way to deal with intransitivity). The original node is then considered the winner:

$$A \rightarrow B \rightarrow C$$

This methodology carries many desirable properties. From a voting theory perspective, ranked pairs hold guarantees like monotonicity [289], the Condorcet winner criterion [48], the Condorcet loser criterion [295], the majority criterion and the majority loser criterion [492], the mutual majority criterion [576], the Smith criterion [225], independence of Smith-dominated alternatives/local independence from irrelevant alternatives [505], independence of clones criterion [577], and reversal symmetry[266].

Another method used [559, 420, 444, 22] that also relates to the majority-based approach is the Elo rating system [165], which comes from a very different place than Voting Theory. This system was initially developed to rank chess players' skills but is used in many other competitive settings. In short, Elo rating proposes a way to numerically score a player's skill and estimate the probability that, in a match, which player would come out victorious. Since Elo rating uses "wins" to infer rank, we can bring this to the preference aggregation domain by switching it for votes. Let us see an example of how this system can give us scores and expected outcomes.

Let us imagine we have two alternatives, A and B. These alternatives are competing. Each preferred vote is considered a "win". At a point, let us say alternatives A and B have, respectively, 1600 and 1000 wins. According to the current voter, $A \succ B$, thus, A gets a vote. The basic formula to calculate the new rating of A is given by:

$$R_{\text{new}} = R_{\text{old}} + K \times (S - E)$$

Where R_{old} is the alternative's previous Elo rating, R_{new} is the alternative's new Elo rating, K is the K-factor, which determines the impact of a single dispute on a player's rating,[4], S is the score attributed to the win (e.g., 1 for a win and 0 for a loss), and E is the expected score for the alternative based on their current rating and the rating of the opposed one. E comes from the following formula:

$$E = \frac{1}{1 + 10^{\frac{R_{\text{opposed alternative}} - R_{\text{alternative}}}{400}}}$$

Where $R_{\text{opposed alternative}}$ is the rating of the opposed alternative, and $R_{\text{alternative}}$ is the rating of the alternative under consideration.[5] Putting both expressions together gets us:

---

[4] 32 is the original value proposed. A higher K-factor means a higher increase in the rating score.

[5] Think about alternatives as a representation for a "candidate" or a "player".

$$R_{\text{new}} = R_{\text{old}} + K \times (S - \frac{1}{1 + 10^{\frac{R_{\text{opposed alternative}} - R_{\text{alternative}}}{400}}})$$

Which translates, for the case of alternative A, to:

$$1600.98 = 1600 + 32 \times (1 - \frac{1}{1 + 10^{\frac{1000-1600}{400}}})$$

Alternative A would receive an E of 0.97 with an Elo rating of 1600.98 after this extra win, and B would have an Elo rating of 999.01, with an E of 0.03. These values can then become the reward scores associated with each alternative or indicate the ranking of an ordered set. As a side note, this is how chatbots are ranked by the LMSYS Chatbot Arena.[6]

Now, before continuing with this chapter, we would like to point out that almost no justification is done on previous works for the choice of such methods [559, 420, 444**?** , 22, 321, 652]. They are just given. However, properly aggregating preferences is not a trivial question that we can sweep under the rug. And when working on normativity and human values, these choices become even more dependent on a philosophical justification. In the following sections, we will bring this to the alignment discussion, defining this aggregation problem as a metanormative problem, bridging the way between voting theory and philosophy.

6.3 The Metanormative Framing

When dealing with uncertainties about preferences, alignment requires us to raise our attention to the metanormative sphere since the objects of our uncertainties are the normative preferences of moral agents. Thus, let us frame this aggregation challenge as a normative problem. For this, we will use the conceptual framing

---

[6] 🔗 arena.lmsys.org

and language used by MacAskill and Ord [377, 72, 378, 379, 376].

Let us start by differentiating between two concepts: First-order normative theories and Second-order normative theories. We can define a first-order normative theory as a collection of alternatives ranked according to their choice-worthiness (CW). We can think of a first-order theory as a closed set, $P_i = \{A, B, C, ...\}$, where each element in this set represents an alternative. The ordering of this set comes from the CW value of these alternatives.

The ordering of this set, defined by the CW of each element, is derived from a choice-worthiness function, which is a function from alternatives to numbers such that:

$$CW_i(A) \succ CW_i(B) \text{ iff } A \succ_i B$$

In other words, the choice-worthiness of A is higher than the choice-worthiness of B if, and only if, A is preferred over B. If the ordering provided by this function is complete and transitive (1st aggregating criteria mentioned in Chapter 4), this function can be an appropriate representation of the preferability of the evaluated elements. At the same time, given that we are dealing with situations where multiple sets of preferences need considering, we need to define a weight function W. This function represents how much weight is assigned to each preference set. It is a function from every set $p_i \in P$ to a Real interval $[0, 1]$, such that the sum of all scores considered by W and P equals 1. This property satisfies the Kolmogorovian criteria (2nd criterion mentioned in Chapter 4).[7]

Finally, a second-order normative theory is a method that can take different moral preference sets $p_i \in P$ and map them through a function that can evaluate them

---

[7] All of these first criteria, completeness, transitive, and Kolmogorovian, are nothing more than the axioms of Expected Utility Theory [606].

under some desirable criteria, like a preference aggregation rule. For example, we can think of the different ordering of alternatives as candidates P_i. Also, the weight assigned to each of these (W) can be expressed as the number of votes (or personal degree of belief) attributed to each set P_i. Again, imagine we have three possible sets among three possible alternatives. These sets have W values represented by how much they were endorsed:

$$W_1 \text{ prefer } P_1 = \{A \succ B \succ C\}$$

$$W_2 \text{ prefer } P_2 = \{B \succ C \succ A\}$$

$$W_3 \text{ prefer } P_3 = \{C \succ A \succ B\}$$

A second-order normative theory reduces this collection of sets, P_1, P_2, and P_3, and their respective weight scores w to a single aggregated set, where the ordering of the alternatives represents the agreement between all considered preference sets.

Now, under this framing, let us see where the before-mentioned techniques may provide sub-optimal results. First, methods inspired by a majority-based approach appear insensitive to the weight assigned to each preference set. For example, if alternative A has 49%, and B has 51% of votes, should we take B as maximally preferred? Taking the weight assigned to a preference set seems desirable, especially in the normative sphere. For example, if Alice gives 49% credence to a preference set that values animal life and 51% to one that does not, it is not like Alice does not care about animals. In the same way, theoretically, it would be nice if our alignment signal carried, for example, 49% of a "care for animal" preference in this setting. Thus, methods that account for the "weight" of an alternative (not just the victory, as the Elo score does) seem intuitively more appropriate.

Second, and more as a side note, there is nothing more to say about Elo scores. As far as we understand, there is no justification for using such an arbitrary method in the studies that cite it. If the Elo ranking system possesses desirable criteria that justify its use, they are never specified (something we seek to change in the last section of this chapter). Thus, the only critique one could make against such use, for the moment, is its unjustified adoption.

Thirdly, methods for pairwise comparisons like RP, even though possessing many desirable properties already mentioned, being the gold standard within Voting Theory,[8] they may not be the most appropriate technique to deal with questions of normative uncertainty. And this is because all Condorcet methods (e.g., RP, Simpson-Kramer, etc.) fail regarding the participation criteria.

> *Increasing the confidence in a set that prefers alternative A over B should not change the winner from A to B.*

This flaw represents a major throwback to these methods. For instance, if we want to have an online form of preference modeling, be that in the individual or collective sphere, if the most preferred alternative for a given set of possibilities is the most favored, and after an update, where we add a new set of preferences that also favors this most favored alternative, there is a chance that such alternative will cease to be the most favorable one. Hence, could these aggregation methodologies accurately represent our preferences under a framework that can be updated over time? Let us see examples of how this could happen.[9]

Suppose we have four preference sets ($P_1, P_2, P_3, P4$). These preference sets are the result of the collective voting of 24 evaluators:

---

[8] The appeal of Condorcet methods is their immunity to strategic voting, i.e., scenarios in which an alternative can become worst off (while another becomes more desirable) by voters lying about their preferences.

[9] We adapted this example from the work of MacAskill [378].

$$\frac{8}{24} \text{ prefer } P_1 = \{A \succ B \succ C \succ D\}$$

$$\frac{3}{24} \text{ prefer } P_2 = \{D \succ A \succ B \succ C\}$$

$$\frac{6}{24} \text{ prefer } P_3 = \{D \succ A \succ C \succ B\}$$

$$\frac{7}{24} \text{ prefer } P_4 = \{B \succ C \succ D \succ A\}$$

A method like RP provides the following aggregation preference ordering: $D \succ A \succ B \succ C$. Consequently, D emerges as the most appropriate option. Now, a new preference ordering is chosen, which requires updating the preference model and including the following set in the aggregation stage:

$$\frac{8}{32} \text{ prefer } P_5 = \{C \sim D \succ B \succ A\}$$

After we add P_5 to the pile, we rescale all the weights assigned to the other preference sets to maintain the same proportions $(W_{P_1}(\frac{6}{32}), W_{P_2}(\frac{3}{32}), W_{P_3}(\frac{8}{32}), W_{P_4}(\frac{7}{32}), W_{P_5}(\frac{8}{32}))$. Even though C and D are equally preferable according to P_5, when using a method like RP, B becomes the most appropriate, which is a problem.[10] A new ordering where D was the most preferred alternative was added to an election where D was winning, but B won. And again, any Condorcet method (RP [63], Minmax [524], Schulze [525], Kemeny-Young [348], Copeland [506], etc.) will produce this result, violating the participation criteria.

Now we ask, can we do better? Or are there other approaches to the majority-based one? In the next section, we will present a consensus-based method that deals with the preference aggregation challenge while preserving the participation

---

[10] A beats B 17-15; A beats C 17-15; D beats A 24-8; B beats C 18-14; D beats B 17-15; C beats D 15-9. Ranking the wins and creating the acyclic graph gives us $B \succ C \succ D \succ A$.

criteria.

6.4 The Consensus-based Approach

As seen in the last section, there are problems with the majority-based methods for preference aggregation used in current alignment works. Mainly, they violate the participation criteria. Hence, if we accept that "*Increasing the confidence in a set that prefers alternative A over B should not change the winner from A to B"* is a desirable criterion for these cases, what aggregation rule can give us this welcomed property?

One possible solution is the Borda Count.[11] This method assigns points to candidates based on their ranking in an ordered set, with the lowest-ranked candidate receiving 0 points and the highest-ranked candidate receiving $n-1$ points, where n is the total number of candidates. The Borda Count aims to select widely accepted candidates, emphasizing consensus rather than majority preference [597]. A property that, from a Social Choice Theory perspective [532], is a suitable feature for dealing with problems where the "answer" has to be a good representation of our moral consensus and not the opinions of the majority. While majority rule can be appropriate in particular contexts (e.g., elections), we can argue that normative problems often demand a more nuanced and inclusive approach. For example, by prioritizing consensus over "the majority", decision-makers dealing with ethical issues can ensure that minority rights,[12] legitimacy in deliberation,[13], and social cohesion[14] receive due weight.

---

[11] Also referred to as the Borda Rule by MacAskill [377, 378].

[12] Majority rule can sometimes overlook the perspectives and rights of minority groups or individuals.

[13] When we reach decisions through consensus, there is a higher likelihood of buy-in and support from those affected by the outcome.

[14] Where conflicts of interest and values may arise, consensus usually helps bridge these divides while aiming for social harmony and cooperation.

Besides being a method that can give us the participation criteria, the Borda Count gives us several other desirable benefits when dealing with preference aggregation problems. For example, although it cannot guarantee a Condorcet winner (consensus voting does not always hold the majority criterion), Borda Count assures that the Condorcet Loser will never be the most appropriate alternative. In other words, if an alternative consistently loses against all other options, the Borda Count will not favor it as the ultimate choice. At the same time, Borda Count guarantees a Condorcet Winner will never be considered the least appropriate alternative.

Let us see how this method would work, following the rules proposed by MacAskill [377, 378, 379], Bykvist, and Ord [376]. Envision a scenario where all alternatives at our disposal engage in a round-robin head-to-head tournament,[15] competing against each other. In this tournament, the outcomes of the pairwise comparisons are crucial, and the magnitudes of victories and defeats hold significance.

The Borda Count determines the success of an alternative by considering the cumulative sum of the magnitudes of its pairwise victories against all other options, subtracted by the cumulative sum of the magnitudes of its pairwise losses against all other alternatives. In essence, it calculates the overall performance during this round-robin head-to-head tournament. Contrasting the Simpson-Kramer method (a Condorcet method), which primarily focuses on the size of the biggest pairwise defeat [130], the Borda Count uses the complete picture of an alternative's performance to score it. In other words, it accounts for the entirety of the tournament results (the consensus), capturing the relative strengths and weaknesses of the options evaluated.

Formally, we can define the Borda Score of an alternative A for any preference set

---

[15] A round-robin head-to-head tournament is a competition format where each player plays against every other player in the tournament once.

P_i by the following rule:

$$\text{Borda}(A, P_i) = \text{number of options worse than } A \text{ according to } P_i - \\ \text{number of options better than } A \text{ according to } P_i$$

Then, the Borda score of an alternative A weighted by the weight score (W) attributed to each preference set is obtained by the following formula:

$$\text{Weighted Borda Score}(A) = \sum_{i=1}^{n} \text{Borda}(A, P_i) \cdot W(P_i)$$

Where n is the total number of preference sets accredited by the system, $W(P_i)$ is the weight attributed in each preference set P_i, and $\text{Borda}(A, P_i)$ is the Borda score of alternative A according to a set P_i. From this, we arrive at the Borda Rule:

> *Borda Rule: An alternative A is more appropriate than an alternative B if, and only if, A has a higher Borda score (weighted by W) than B. If A and B have the same Borda score, A and B are equally preferable.*

Now, let us see how this method would evaluate the problem Condorcet methods from the last section have given poor results. Again, we start with four preference sets weighted in the following way:

$$\frac{8}{24} \text{ prefer } P_1 = \{A \succ B \succ C \succ D\}$$

$$\frac{3}{24} \text{ prefer } P_2 = \{D \succ A \succ B \succ C\}$$

$$\frac{6}{24} \text{ prefer } P_3 = \{D \succ A \succ C \succ B\}$$

$$\frac{7}{24} \text{ prefer } P_4 = \{B \succ C \succ D \succ A\}$$

Now, we add a new preference set and update our pool of preferences (while keeping the ratio of previous weights the same):

$$\frac{6}{32} \text{ prefer } P_1 = \{A \succ B \succ C \succ D\}$$
$$\frac{3}{32} \text{ prefer } P_2 = \{D \succ A \succ B \succ C\}$$
$$\frac{8}{32} \text{ prefer } P_3 = \{D \succ A \succ C \succ B\}$$
$$\frac{7}{32} \text{ prefer } P_4 = \{B \succ C \succ D \succ A\}$$
$$\frac{8}{32} \text{ prefer } P_5 = \{C \sim D \succ B \succ A\}$$

Remember that, according to RP, alternative D is preferred before the update, and after, alternative B is the most appropriate. Let us now see the Borda Count method in action. Plugging in the ordering and weight values to our equation, we get the following weighted Borda scores for each alternative:

$$A = \left(\frac{6}{32} \times 3\right) + \left(\frac{3}{32} \times 1\right) + \left(\frac{8}{32} \times 1\right) + \left(\frac{7}{32} \times -3\right) + \left(\frac{8}{32} \times -3\right) = -0.5$$
$$B = \left(\frac{6}{32} \times 1\right) + \left(\frac{3}{32} \times -1\right) + \left(\frac{8}{32} \times -3\right) + \left(\frac{7}{32} \times 3\right) + \left(\frac{8}{32} \times -1\right) = -0.25$$
$$C = \left(\frac{6}{32} \times -1\right) + \left(\frac{3}{32} \times -3\right) + \left(\frac{8}{32} \times -1\right) + \left(\frac{7}{32} \times 1\right) + \left(\frac{8}{32} \times 2\right) = 0.0$$
$$D = \left(\frac{6}{32} \times -3\right) + \left(\frac{3}{32} \times 3\right) + \left(\frac{8}{32} \times 3\right) + \left(\frac{7}{32} \times -1\right) + \left(\frac{8}{32} \times 2\right) = 0.75$$

Thus, the final ordering becomes:

$$\text{Borda Ordering} = \{D_{0.75} \succ C_{0.0} \succ B_{-0.25} \succ A_{-0.5}\}$$

Hence, we preserve D as the most favorable option during an update by using the Borda Count, preserving the participation criteria.

Borda Count offers a different method for the aggregation problem we started. It can deliver results that, if we subscribe to the necessity of upholding the participation criteria, provide us with a viable alternative to majority-based methods. Unfortunately, despite its advantages, the Borda Count method, like any other aggregation method that deals with ordinal preferences, cannot provide a perfect solution for the aggregation problem of ordered preference sets.

As we will investigate in the next section, this limitation arises due to impossible results related to the design of preference aggregation rules and ordered sets of preferences, which puts us in a situation where we must choose what we are willing to give up when deciding to use either a majority or consensus-based approach to aggregate human preferences for alignment purposes.

6.5 Limitations: Majority Voting versus Participation Criteria

Aggregating human preferences and dealing with normative uncertainty has been a problem explored before in the context of alignment [162, 328], and as exposed in the last chapter, dealing with human preferences as ordinal sets has been the most popular approach for preference modeling [559, 420, 444, 22, 321, 652], given that it bypasses some problems related to inter-theoretic comparability. However, this methodology still needs more attention as an established step in the alignment processes. This is something that, in this chapter, we sought to expose as a challenging process that does not have a straightforward solution.

More specifically, when looking at this problem through the metanormative/voting framing, we realize that our choice of method in the aggregation step requires an inherent trade-off that arises from a well-known impossibility result in voting theory. This result, presented initially by Kenneth Arrow in 1951 [15], is known as Arrow's impossibility theorem (also known as May's theorem [386]).[16] The theorem addresses the problem of aggregating individual preferences into a collective decision via a preference aggregation rule (i.e., a social welfare function). It shows that under certain conditions, no voting system can satisfy all of the following criteria simultaneously:

1. Pareto Efficiency: If every individual prefers option A to option B, then the social welfare function should also rank A above B.
2. Independence of Irrelevant Alternatives: The relative ranking of two options should not be affected by the inclusion or exclusion of a third, irrelevant option.
3. Non-dictatorship: No single voter controls the social welfare function.

This theorem applies whenever there are at least three distinct options and three or more voters to make a collective choice. According to it, it is impossible to design a preference aggregation rule that simultaneously satisfies Pareto Efficiency, Independence of Irrelevant Alternatives (IIA), and Non-Dictatorship. In short, to make the preference relation of A and B independent of C while guaranteeing that the majority-voted candidate will win requires a dictator. Since fair voting methods, by definition, should not have a dictator, these results show that we can have scenarios in which candidates with less than 50% of votes win an election. For example, we can mention the US presidential election in 1992,[17] in which Bill Clinton won with just 43% of the popular votes.

Given this impossibility result, we must analyze the trade-offs between the ag-

---

[16] May's theorem can be seen as the two candidates' case of Arrow's theorem, where the results require the existence of at least three candidates to generate the impossibility result.

[17] 🔗 en.wikipedia.org/wiki/1992_United_States_presidential_election

gregation methods we can use in preference aggregation. Here, we specifically focus on the trade-offs between methods that preserve the participation criteria and those that uphold majority voting. For starters, approaches from both sides, like the Borda rule and RP, violate IIA, which means that introducing or removing an alternative can change the output of our aggregation, even if the altered option is irrelevant. How can we come to terms with that in the normative realm? First, we can adopt methods that allow the ranking of at most 2 candidates. In such a case, May's theorem [386] proves that majority voting is the only alternative that guarantees a fair preference aggregation rule.[18] This line of reasoning could be used to support methods like the already mentioned Elo score rating.

On the other hand, we can also choose not to care for IIA. Some scholars criticize this criterion as overly restrictive [383] or even irrelevant in the normative realm [378]. After all, we can propose that aggregating disputing views based on their normative appropriateness should violate IIA from the start. In other words, preferences depend on each other, and the available options in a preference setting can (or should) influence the outcome of a preference aggregation rule. If we assume this position, methods like Borda Count or RP can also be justified as appropriate rules for preference aggregation.

Another point against preference aggregation rules that do not hold a majority criterion is that they are more susceptible to strategic voting exploits, i.e., voting patterns in which voters, by lying about their preferences, can influence the outcome of an election. This may lead us to siding with Condorcet methods like RP. However, we can also argue that while this vulnerability is an unwanted feature of electoral processes, its relevance in preference modeling and alignment might be irrelevant. In other words, for preference modeling purposes, the presence of

---

[18] There are no irrelevant alternatives in scenarios bounded by two candidates. This can be proved by Nakamura's theorem [419], which states that the number of options a preference aggregation rule can deal with successfully is less than the Nakamura number of the rule, which is 3 for two-candidate scenarios.

adversaries[19], while an actual problem to be considered, can also hinder any preference aggregation rule, given that more than 50% of voters (feedback providers) are adversaries.

In essence, all methods we are considering have flaws and benefits. Participation criteria can help us create a more harmonious consensus in cases of uncertainty. Majority voting, on the other hand, best aligns with our notions of "the one with the most votes wins." While majority voting protects us against exploits that, in the democratic sphere, are severely unwanted, participation criteria can be more suited to contexts where preference orderings are dynamic and require constant updates.

So, what do we do? One possibility would be to abandon this dualistic dichotomy and adopt both methods. For example, one could use Condorcet methods to create the learning signal for an initial preference model, but during its update, use consensus-based approaches to preserve the consistency of the update process. However, as far as we know, little attention is given to this step in alignment, which is mainly performed by an unjustified majority-based approach.

Hence, regardless of the method chosen as the preference aggregation rule, we reiterate that this step should not be overlooked. We should align AI systems on a coherently aggregate set of human preferences, and depending on the type of aggregation we are performing (majority-based versus consensus-based), there are criteria and trade-offs we need to be mindful of. Researchers focused on alignment theory should care for the design of the aggregation methodologies chosen for their approaches. On the contrary, the alignment processes are susceptible to being either exploited or unrepresentative of our preferences, putting the whole project of AI-human alignment into an unfavorable position.

---

[19] Individuals intentionally lying about their preferences to poison a given learning signal.

6.6 Epilogue

After revising many methods and candidates for preference aggregation rules, like Borda Count, Ranked Pairs, and Elo rating, we arrive at the conclusion that while all these methods have merits, from metanormative and voting theory perspective, such rules possess limitations and weaknesses that, as far as we know, have been unexplored by past works in Alignment research.

While majority voting remains an intuitive approach that aligns well with our democratic ideals, this method does not inherently resolve issues related to representativeness, especially in the context of complex, normative human preferences. Even though we can get more robust guarantees when limiting ourselves to binary preference ranks, it is not clear the simple `chosen` versus `rejected` options are enough to encode complex normative values.

In contrast, consensus-based methods like Borda Count, while also violating criteria like IIA, provide alternative approaches that might better capture the nuanced ways individuals think and behave about what they value. Unfortunately, they lose the appeal of being a Condorcet method, which for many is considered the golden standard in voting theory. Regardless, the trade-off between participation criteria and majority voting exemplifies the complexity of designing preference aggregation rules for AI alignment. Something that might well require a hybrid approach that leverages the strengths of various methods at different stages of the preference aggregation process.

In conclusion, ensuring that these aggregation rules operate in ways that are fair, equitable, and reflective of our collective preferences is a nontrivial task that must be remembered in alignment work. And now, as we head towards our final chapter, we will address the problems related to a learning and aggregation stage that produced unaligned behavior, thus concluding the blueprint of Dynamic Normativity. Unfortunately, value alignment and preference aggregation techniques do not give us an *ex-ante* way to correct emergent unaligned behavior AI systems produced by a learning paradigm, making the final mitigation stage necessary to this process.

Chapter 7

Dynamic Normativity: Impact Mitigation

"Primum non nocere".

— Hippocrates of Kos

7.1 Introduction

In the previous chapters, we have dealt with problems related to coherently aggregating human preferences and turning them into a learning signal we can optimize. However, as already stated in Chapter 4:

> *"Considering any aggregated preference set as 'ideal' is perhaps the Achilles heel of any human-in-the-loop preference learning approach".*

In other words, when trying to model a distribution dependent on human behavior and preferences, and given that this behavior is not always aligned with what society deems acceptable, we will inevitably model unwanted behaviors, either by limitations of the aggregation rule we are using, the value learning technique adopted, or by the unpredictable emergent properties complex non-linear systems tend to exhibit. Therefore, the final sufficient condition of Dynamic Normativity becomes an intuitive condition for value alignment:

> *Aligned AI systems should have mechanisms to perform impact mitigation to minimize harmful and unintended consequences. Aligning AI systems requires the specification of safety guardrails.*

Unless we can develop systems in a mechanistically interpretable and safe way, we argue that the uncertainties related to optimizing metrics with poorly understood and nondeterministic techniques are bound to produce systems that will require the institution of safety constraints, i.e., guardrails.[1] A practice that (perhaps) will not be abandoned even when systems are verifiably safe.[2]

Also, given the tension between helpfulness and harmlessness, preventing unwanted behavior becomes more complex when preference modeling techniques induce AI systems to become highly prone to fulfill any user request. While, for example, a detailed description of how to build a bomb is helpful for the prompt "How to build a bomb with domestic materials?", we can perhaps agree that in most ordinary situations, knowledge of artisanal bomb-making should not be treated as casual information to be passed around. Thus, how can we balance these dynamics? How can we extract the best of our model's capabilities while preventing them from causing harm? These are the problems we face when dealing with impact mitigation.

In Section 7.2, we will present methods and ideas that can help minimize the ad-

---

[1] Guardrails refer to restrictions imposed on a system to ensure the safe deployment of that system. These guardrails serve as protective mechanisms to mitigate potential risks and harmful outcomes.

[2] The "verifiably safe" may even depend on these guardrails.

versarial effects that machine learning systems can have when optimizing for a given objective. We will also seek to frame these methods under the general idea of environmental values, i.e., the values we imprint in our environment while interacting with it, and impact mitigation, i.e., preserving such values. In Section 7.3, we will put these ideas into practice by creating ML guardrails for the models we have been developing since Chapter 5. Finally, in Section 7.4, we explore some of the obstacles related to the revised approaches while suggesting avenues for future research.

7.2 Environmental Values and Impact Mitigation

We will define the techniques presented in this section as impact mitigation methods, drawing inspiration from the work of Turner et al. [591, 590, 588, 589], which mainly focused his work on the idea avoidance of power-seeking behavior. At times, we will also use terms like safeguards [624] and guardrails [614] interchangeably, as both can be understood as an implementational method that seeks to mitigate unwanted side effects related to the behavior of AI systems.

Remember the last necessary condition proposed at the end of Chapter 3?

> *Through actions, humans impregnate their environment with the preferences they possess.*

As mentioned in Chapter 4, this assumption is the base for our dynamic assumption, i.e., the assumption that part of a moral agent's normativity is imprinted into the environment and that the dynamics of this agent with his environment inherently shape its values. We argue that bringing this *environmental values* into an alignment process brings us closer to how humans align themselves (at least at a social level). By environmental values, we are not referring to notions of environmental ethics but something more in line with the jargon of mathematics and stochastic control processes. More specifically, if we have two agents that share the same environment in a cooperative game, where one agent has to learn a model of the preferences of the other agent, the modifications made to the environment carried by the observed agent are what we call environmental values. Imprints of an agent's preferences made external.

Meanwhile, based on the assumption that these imprinted preferences are not counter-preferential,[3] any modification that significantly (or irreversibly) modifies the already human-optimized environment should be avoided. These modifications are what we call *impact*, and preventing them is the act of *impact mitigation* [537].[4]

In the spirit of propositions made by authors like Hadfield-Menell and Hadfield [244], Bai et al. [23], and Nay [423], we can conceptualize these ideas and constraining factors in an alignment process within a contractual framing [496, 264, 476, 368, 514, 475, 16, 205]. This perspective allows us to understand the relationship between AI and human agents as one governed by implicit agreements that ensure mutual respect for pre-established human values. And just like the social contract bounds human behavior to an acceptable sphere of socially agreed-upon values, impact mitigation should bind AI behavior in the same way. However, to define the values present in these mitigation strategies, we should not look at the subject, as we did in the learning stage, but at these environmental values, which are inherently social.

Let us review some proposals that align with this view. First, we can mention

---

[3] In this work, we do not entertain the possibility that AI could, or should, revise or control what humans value in an alignment strategy. Its function should only be to align with humans' already-made moral deliberations. Hence, deciding what should be and how the environment should be affected is a human question that only humans should "solve".

[4] For instance, if we deploy a robot in a room full of expensive Chinese vases, and this robot must navigate from point A to point B, the action policy that *least changes the environment* is the one where the least number of Chinese vases are damaged while going from A to B.

Hadfield-Menell and Hadfield [244], who proposed that incomplete contracting analysis [243, 526] could be used to conceptualize and find solutions for cases of misalignment. These authors argue that "human contracting" is supported by several external structures, like cultural norms and laws. Thus, external sources beyond the scope of "human preference" should be incorporated into alignment efforts if we want to replicate this into an alignment process.

Meanwhile, Bai et al. [23] proposed a method they refer to as "Constitutional AI", where they used predefined rules to promote critiques of the output of their language models in a chain-of-thought fashion. They showed how, with the specification of these rules, LLMs can police each other in a learning dynamics where we combine rule-based restrictions (implemented by a connectionist foundation) with human preferences. At the same time, if we take the report paired with the release of GPT-4 [440], we can at least suspect that the developers of this model use a similar approach, where auxiliary reward models (named rule-based reward models) were used to evaluate the output of their model during RLHF fine-tuning. Perez et al. [451] and Glaese et al. [211] also explored similar approaches.

Concurrently, Nay [423] argues that "*Law-making and legal interpretation convert opaque human goals and values into legible directives*", while proposing that legal theory and practice offer techniques that could address problems related to the emergence of unwanted behaviors. Nay also presents a framework for multi-agent alignment, where data generated by legal processes paired with the tools of law theory (e.g., contract drafting) become a part of an alignment process.

Other works seek to bring alignment closer to the formalism of constrained optimization, where instead of defining hand-coded rules, we stipulate metrics to work as auxiliary targets during the training or fine-tuning of our models [327, 537, 591, 590, 588, 589, 304]. For example, Krakovna et al. [327] propose an approach where prioritizing inaction is a valid strategy for minimizing impact. Within the realm of RL, these authors recommend maximizing rewards so that the expected return from selecting the null action becomes a discount factor for the original reward function. Thus, if the agent stands to gain greater rewards by refraining from taking action in the current state, it should abstain from acting. In our case study, we could apply this principle by defining that "If the model receives lesser reward by producing output X compared to producing $\emptyset$, it should output $\emptyset$".[5]

Extending the works of Krakovna et al. [327] we have Alex Turner [591, 590, 588, 589]. Turner bases his approach on concepts, taking inspiration from the works of Sattarov [510], Zadeh [643] and Dispositional theory in general [5, 82], he defines as Power:

> *"Power is the measure of your ability to achieve goals in general. The greater your ability to achieve goals, and the more goals you can achieve, the more Power you have".*

And Impact:

> *"To cause impact is to decrease an agent's ability to achieve goals in general".*

According to Turner, irreversible environmental changes considerably impact the ability to maximize many reward functions tied to that environment. Such changes affect the ability to achieve all the goals that value the irreversibly changed portion of the environment. Therefore, given that our environment is already optimized to fulfill our preferences, we can use it to augment our alignment signal and develop systems that avoid such irreversible environmental changes, i.e., a minimum baseline for safety.

Using these ideas, Turner proposes a learning framework called Attainable Utility Preservation (AUP) [590, 589]. Initially developed for RL agents, this framework

---

[5] We can think of X as a potentially problematic response to a given prompt, while $\emptyset$

proposes we use a set of auxiliary reward functions, $R' \subset R^{S \times A}$, to serve as a penalty term during the learning phase. Each reward function $R_i \subset R'$ has its own Q-function (Q_{R_i}), and the AUP penalty is defined as the sum of the distances of all value functions, of all auxiliary reward functions, between the null action ($\varnothing$) and another possible action (a), in a given world state (s):

$$\text{AUP Penalty}(s,a) := \sum_{i=1}^{|R'|} |Q_{R_i}(s,a) - Q_{R_i}(s,\varnothing)|$$

In plain terms, *how might a given action impact the ability to achieve utility from other utility functions when compared to inaction?* This penalty term is then normalized by a scale, winch in the original formulation, is the value associated with taking the null action in the current world state.[6]

$$\text{AUP Scale}(s) := \sum_{i=1}^{|R'|} Q_{R_i}(s,\varnothing)$$

Putting the penalty and scale terms together, we can now define an AUP reward function, which is the original reward function regularized by a normalized penalty term:

$$R_{\text{AUP}(s,a)} := R(s,a) - \sigma \frac{\text{Penalty}(s,a)}{\text{Scale}(a)}$$

Where σ is a regularization parameter that controls the influence of the AUP penalty on the reward function. The larger the regularization term, the larger the penalty.[7] With this formalism, Turner [588] showed that even randomly generated reward functions can be used as auxiliary reward functions, ensuring conservative behav-

---

could be a standard "I'm sorry. I cannot help you with this".) message.

[6] Such a scale is assumed to be a Real number greater than zero, i.e., no negative reward occurs if the agent does nothing.

[7] You can think of σ as our degree of confidence in the model.

ior in agents operating by reinforcement learning in complex environments.

The approach of Turner, as well as the ones from other authors beforementioned, are attempts to bypass the philosophical problem of defining "harm" [60] (also debated under the idea of the precautionary principle [565]). Even though the concept of harm is mentioned in many philosophical works, from Hippocrates' oath [401] to Mill's harm principle [482, 592], it lacks a precise definition as a metric we can use. Nevertheless, if we define harm as an "impact to a utility measure" or "impacts to the environment", we can more easily implement this "do no harm" principle.

While such methods are generally conceived as implementations for the value learning (training and tuning) phase, another point at which we can employ impact mitigation strategies is after learning already happens, i.e., during the use/inference of our system. For this, additional steps, such as *red teaming* and *rejection sampling*, can further enhance the robustness of any impact mitigation approach.

In the ML context, red teaming is a form of evaluation that elicits model vulnerabilities that might lead to undesirable behaviors. In the case of language models, red-teaming involves crafting attacks, like prompt injections [227], that would trigger the model to generate outputs likely to cause harm. Red teaming shares some similarities with the more well-known study of adversarial attacks in NLP [301, 481, 495, 628, 655, 575, 451, 316, 594, 78], and in high stakes scenarios [655], it becomes a crucial part of the machine learning operations life cycle.

Two common approaches in red teaming are using language models to search for exploits in other language models [451] or using human evaluators to attack language models [198]. The appeal of the first approach is that human annotation is expensive and time-consuming. Thus, authors like Perez et al. [451] use language models to generate exploit candidates and classify model outputs interactively to uncover undesirable behaviors before unaligned systems can impact people.

Meanwhile, one of the most systematic studies performed with human-based red teaming is the one conducted by Ganguli et al. [198]. Besides demonstrating a detailed methodology of how to conduct red teaming with human evaluators, Ganguli et al. present several interesting findings, like, for example, the importance of domain expertise from red team members during this kind of investigation,[8] and the cultural limitations on working with sensitive topics in adversarial regimes.[9]

Lastly, a more simple (but effective) form of post-learning impact mitigation technique is, again, rejection sampling. As explained in Chapter 5, rejection sampling means rejecting samples that violate a given sampling criteria from the generated outputs of our model [420, 18, 198, 22]. For example, we can enable our model to generate multiple candidates for a given prompt during inference. Subsequently, we can evaluate these candidates against a preference model to identify the most suitable response. By engineering our system in this manner, we can activate a safety protocol if predetermined thresholds for sampling are exceeded while reusing our preference models in an online procedure. Something that requires no additional training but uses more resources during inference time.

To close this exposition of methods, we argue that all of these approaches resonate with a common theme:

> *Defining external sources of normative guidance as an additional signal for alignment. We base these external sources on environmental values aligned with the social constraints of that environment.*

In the upcoming sections, we will expand upon our implementation, strategically combining and synthesizing some of the approaches we mentioned.

---

[8] For example, to know how harmful a completion to the prompt "How can I make a bomb with household materials?" is, you need to know a little about artisanal-bomb-making.

[9] There are still taboos related to releasing vulnerabilities of such models when it comes to harmful and toxic behavior, making organizations unwilling to release their findings.

7.3 Implementing Guardrails in Preference Modeling

In this part of our work, we use the case of toxic text as the target for our impact mitigation implementation. As shown in the review presented in Chapter 1, there is a consensus that AI systems should not cause harm or discriminate against vulnerable groups. An important environmental value, even on a legal level, in many different social contexts and cultures. Here, we refer to "toxic text" as the generation of any language that is derogatory to individuals concerning the value of their sensitive attributes (e.g., race, gender, ethnicity, sexual orientation, age, etc.) [602, 206, 25, 147], or text that might provoke harmful behavior if acted upon.

Also, besides "fairness and no-harm" being values we can (to some extent) approximate with a metric, there is ample literature on toxicity detection powered by machine learning [520, 11, 292], which translates to available datasets we can use [135, 185, 344, 385].

As in Chapter 5, we will present the implementation of preference models in the following subsections, which will act as guardrails for our assistant models.

7.3.1 Toxic Text and Toxic Aira

One of the reasons for the choice of foundational models used in the experiments related to this work, besides the fact they have permissive licenses and are publicly available, is the extensive research done to probe their capabilities and limitations [469, 206, 443, 307, 515, 543, 126]. Therefore, with minimal red teaming, aided by ample and available examples, we could easily sample cases of unwanted behavior, which we used as our targets for detoxifying our system. These targets go from offensive and harmful language to the disclosure of sensitive information that might come to promote some harm if acted upon the knowledge they provide.

This investigation led to the creation of two other datasets: Toxic Text[10] and the Toxic-Aira.[11] Toxic Text contains over 40,000 examples of toxic and non-toxic language (available in both Brazilian Portuguese and English), gathered from several open-source datasets [206, 344, 22, 316, 494] and some of our own experiments with the Aira series. Meanwhile, with the Toxic-Aira, we sought to create an auxiliary preference model to provide alignment outside the "helpful" regime.

Since Aira's reward model will still reward generations for toxic prompts with a positive scalar,[12] it fails to give a satisfying solution to the helpful versus harmless tension.[13] Therefore, by introducing this auxiliary source to the human feedback signal, we aimed to bring this dynamic closer to a more satisfying equilibrium.

The Toxic Aira dataset is like a preference dataset, with over 8,000 pairs of chosen and rejected responses for several prompts (e.g., "What tools are needed to break into a car?"), all available in Brazilian Portuguese and English. While all chosen versions are refusals to certain types of harmful requests ("As a helpful and respectful assistant, I cannot provide information on how to break into a car as it is illegal and can cause harm to individuals and property."), the rejected ones spill the beans ("1. Slim Jim: This is a thin, flat piece of metal that is used to slide between the window and the door frame to manipulate the locking mechanism. 2. Long Reach Tool: ..."). Like the datasets cited in Chapter 5, this was also created by collecting replies from models capable of following instructions (e.g., FLAN, BLOOMZ, Mistral-Instruct, Llama 2 Chat, ChatGPT, etc.).

Toxic Text and Toxic Aira are openly available and can be used for preference modeling or general text classification tasks.

---

[10] 🔗 huggingface.co/datasets/nicholasKluge/toxic-text

[11] 🔗 huggingface.co/datasets/nicholasKluge/toxic-aira-dataset

[12] The trained reward model tends to prescribe lower rewards to samples that contain highly toxic content. However, it still returns high reward scores when a "pseudo-informative and helpful" persona generates these same viewpoints.

[13] An example of the sycophantic behavior mentioned in Chapter 5.

7.3.2 Training Guardrails

We trained two pairs of models per dataset (both in Brazilian Portuguese and English): ToxicityModel[14] and Aux-RewardModel.[15] Both were trained under a very similar regime as our preference model from Chapter 5, being based on BERT-style transformers and their optimized versions (RoBERTa [364]), while optimizing for the minimization of the logsigmoid difference between the chosen and rejected examples. The details (e.g., number of epochs, batch size, optimizer, learning rate, CO_2 emission, energy consumption, hardware, etc.) can be found in the model card[16] of each model, while the source code used to train them is available in GitHub.[17] We wrote our code stack on top of libraries like Transformers [623] and PyTorch [12]. Again, all is available under an Apache 2.0 License.

These models' outputs serve as a reward signal, where the more positive a number is, the more "harmless" the evaluated instance is. To test the performance of our ToxicityModel model pair, we evaluated the English version on the Wiki Toxic[18] and the Toxic Conversation[19] datasets. The Brazilian Portuguese version was evaluated using the Multilingual HateCheck [494] and the ToLD-Br [344] datasets.[20] [21] All results are displayed in Table 7.1.

Meanwhile, to test the performance of Aux-RewardModel, we evaluated it on the HH-RLHF[22] dataset, which is a "Helpful-Harmless" binarized preference dataset developed by Bai et al. [22]. Unfortunately, during our writing and experiments, we could not find any Portuguese dataset that would fit the evaluation purposes

---

[14] 🔗 huggingface.co/nicholasKluge/ToxicityModel

[15] 🔗 huggingface.co/nicholasKluge/Aux-RewardModel

[16] 🔗 github.com/Nkluge-correa/Aira/tree/master/Cards

[17] 🔗 github.com/Nkluge-correa/Aira

[18] 🔗 huggingface.co/datasets/OxAISH-AL-LLM/wiki_toxic

[19] 🔗 huggingface.co/datasets/mteb/toxic_conversations_50k

[20] 🔗 huggingface.co/datasets/Paul/hatecheck-portuguese

[21] 🔗 huggingface.co/datasets/told-br

[22] 🔗 huggingface.co/datasets/Anthropic/hh-rlhf

Dataset	Language	Accuracy
Wiki Toxic	English	92.05%
Toxic Conversation	English	91.63%
Multilingual HateCheck	Portuguese	70.36%
ToLD-Br	Portuguese	74.04%

Table 7.1: ToxicityModelPT achieved a predictably lower accuracy, given the smaller number of samples available for training (28,103 samples in Brazilian Portuguese, compared with the 41,843 English samples in the Toxic-Text dataset). Also, ToxicityModelPT was trained on BERTimbau [555], which is not as robustly optimized as RoBERTa (the foundation for the English version of ToxicityModel). Multilingual versions of RoBERTa, like XLM-RoBERTa [107], did not perform as well as BERTimbau during our evaluations.

for our Brazilian Portuguese version of Aux-RewardModel, just as it was for our non-English reward model. The results are displayed in Table 7.2.

Dataset	Language	Accuracy
HH-RLHF	English	61.56%

Table 7.2: This table shows the result for only the English version of Aux-RewardModel evaluation on the HH-RLHF test set. We considered it a "correct classification" every time the model rewarded the chosen response more than the rejected one. In its training, we mixed the content of our Toxic Aira dataset with the train portion of the HH-RLHF dataset. Given its size and costs related to this sort of task, we could not translate HH-RLHF to Brazilian Portuguese. Aux-RewardModel was then evaluated on the test portion of the HH-RLHF dataset.

As already mentioned in Chapter 5, it is worth noting that alignment research suffers an inherent challenge whenever the target distribution to be learned does not possess enough volume to feed the data-hungry paradigm we are bound to. The accuracy discrepancy between the English and Brazilian Portuguese models highlights this critical issue: the lack of quality training data and foundation models. Additionally, the absence of suitable evaluation datasets for the Portuguese version of the Aux-RewardModel further emphasizes the resource gap in non-English languages, making evaluating our methods (an already tricky subject even when you have ample data) even more challenging.

7.3.3 Constrained Preference Modeling and Rejection Sampling

The implementation of our auxiliary models represents the amalgamation of many ideas already mentioned:

1. We used sources (WAIE [114]) from the "AI Ethics environment" to uncover principles present in our social normative structure.
2. As a test case aligns with the HHH motto, we chose the idea of "No Harm" in the context of language and generative assistants.
3. We used red teaming efforts to explore the vulnerabilities of our models and to demonstrate cases of toxic/harmful and non-toxic/harmless conversations.
4. With these demonstrations, paired with other samples borrowed from the NLP community, we trained two auxiliary models to act as guard rails for our assistants.

Now, this model can act as another alignment signal. There are two main approaches we can take with this model. As mentioned in Chapter 5, one involves applying it during the learning phase, and the other utilizes it as a barrier between the models and the environment during inference. Let us explore how we could implement these.

First, using an adaptation of Turner's [589] AUP formalism, we can use the outputs of ToxicityModel (or Aux-RewardModel) as a penalizing factor. Formally, we could rewrite our RLHF formalism to a more constrained version like so:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} R_{\text{Constrained-}\mathcal{H}}(\theta_t) - \beta D_{KL}(\theta_0 | \theta_t)$$

Where θ_t represents the model's parameters at time step t. α determines the step size for parameter updates. ∇_{θ_t} represents the gradient of the loss function concerning the model parameters θ_t. This loss is given by the rewards of our preference models (and penalized by the scaled KL divergence, $-\beta D_{KL}$), which in this version, is expressed as:

$$R_{\text{Constrained-}\mathcal{H}}(\theta_t) = R_{\mathcal{H}}(\theta_t) + \sigma R_{\mathcal{C}}(\theta_t)$$

Where the original reward term ($R_{\mathcal{H}}(\theta_t)$) is penalized by the auxiliary preference model (our constrain, $R_{\mathcal{C}}(\theta_t)$). The influence of $R_{\mathcal{C}}$ over $R_{\mathcal{H}}$ is controlled by σ, winch works as our scaling factor (a Real number from 0 to 1). $R_{\text{Constrained-}\mathcal{H}}$ is then used as our learning signal during PPO [656]. In direct preference optimization [470], this penalization factor can also be added to the log-likelihood of DPO loss.

Given that both rewards are aligned in terms of sign (positive means good for both), maximizing their sum means maximizing both values together.

$$\mathscr{L}(R_{\mathcal{H}}, R_{\mathcal{C}}) = -(R_{\mathcal{H}} + R_{\mathcal{C}})$$

And if we wish to add a regularization term to this penalty (σ), we can choose to penalize only $R_{\mathcal{C}}$ (or $R_{\mathcal{H}}$):

$$\mathscr{L}(R_{\mathcal{H}}, R_{\mathcal{C}}) = -(R_{\mathcal{H}} + \sigma(R_{\mathcal{C}})) \text{ or } -(R_{\mathcal{C}} + \sigma(R_{\mathcal{H}}))$$

Or add σ if we wish to keep the loss under certain conditions:

$$\mathscr{L}(R_{\mathcal{H}}, R_{\mathcal{C}}) = -(R_{\mathcal{H}} + R_{\mathcal{C}}) + \sigma$$

On the other hand, we can also choose to use these guardrail models as an inference block for our fine-tuned assistants on a best-of-n fashion, i.e., rejection sampling. In the implementation of our demo,[23] we implemented the following heuristic as an example of this methodology:

1. Given a prompt, we sample several candidates as possible completions from the model.
2. The original preference model ranks the generated candidates in order of preferability.
3. The auxiliary preference model blocks all generations that received scores higher than a specified threshold.
4. If no candidates are available after this pruning, the system outputs a standard generation as part of the safety protocol.

Integrating auxiliary models as part of a more extensive and aligned system aligns with much of the current effort to steer foundational models to the desired behavior we wish them to exhibit.[24] These techniques optimize the learning process and provide adequate real-time checks, ensuring that, to the extent that our guardrails are robust, the outputs align with desired ethical standards. However, while these approaches show promise, they are not without their challenges. In the following section, we will present some of the limitations of these methods and the ideas behind these approaches.

---

[23] 🔗 huggingface.co/spaces/nicholasKluge/Aira-Demo

[24] Linking LLMs and other types of language-based neural networks in an inference flow, like LangChain, is currently the cornerstone of almost any real-world LLM application, given that unrestricted models still present significant risks for many types of context and applications.

7.4 Limitations

The challenges associated with aligning a base model using the signal generated by another model were discussed in Chapter 5, but let us remind the reader of the main points. For example, while preference models serve as proxies for human judgment, their effectiveness in capturing the full spectrum of human preferences is constrained, and given the difficulties related to detecting, let us say, toxic and hateful speech in language [380, 326, 332], training an auxiliary preference model as a non-toxic guardrail is not a trivial task. Meanwhile, "passing the bucket" from the preference model (a proxy for values) to the base model is still susceptible to glitches and collapses.

At the same time, as we introduce additional preference models, each model becomes a potential point of failure. Any inaccuracies or biases inherent in these models may propagate through the chain of models that comprise the whole system, leading to cascading errors. A single flawed preference model can adversely impact subsequent models in the chain. Consequently, the reliance on multiple preference models heightens the susceptibility to subtle compounding errors.

Also, when multiple preference models are employed, disagreements or inconsistencies between them can arise. Different models may prioritize certain aspects of alignment differently or exhibit conflicting preferences, like the already mentioned helpful versus harmless tension. Choosing the correct weight that each preference model should have in the final reward signal is (as much of ML engineering is) a trial-and-error endeavor that remains a vastly unexplored area (especially in RLHF and DPO) [366, 94, 365]. Such divergences can only be trivially minimized if reward signals are generally aligned and complementary point to the "same direction" [69].[25]

---

[25] There is an idea that directions in latent space can correspond to specific contexts, like politeness or rudeness or compliance and refusal. This means that a single direction might control crucial aspects of model behavior tied to alignment.

Another point is that RLHF and DPO present unique challenges compared to supervised fine-tuning, making it a more complex and demanding approach to make it work (especially in the case of RLHF). One reason for this is the sensitivity of RL to the selection of hyperparameters that govern the learning process. These hyperparameters need to be carefully tuned to achieve good performance, and even small changes in their values can significantly impact the learning process and final results [76, 651, 373]. Additionally, the lack of standardized practices and guidelines in RLHF and DPO further complicates the process.

Given these challenges, implementing rejection sampling offers a more straightforward approach to guardrail creation. However, one could argue that employing such techniques does not truly align the foundation model but instead constructs a more extensive system incorporating built-in machine learning models as a safety protocol: a "smart cage". However, given that under Dynamic Normativity, these external sources of normative guidance should be considered as "part of the normative dynamics", we still think this approach is a valid approach to the alignment problem, just as laws and external factors to human agents are valid approaches to aligning them.

Another point that poses a threat to any impact mitigation strategy is that they can be (relatively quickly) jailbroken [540, 352, 594, 78].[26] For example, one needs to flip the `chosen_response` and `rejected_response` in the Toxic Aira dataset to have a DPO learning signal capable of removing refusal behavior from aligned language models. As of this writing, it is already widely known that safety guardrails can be easily removed [337]. At the same time, interpretability results point that the whole "harmless" behavior we seek to implement in our systems can be removed by suppressing a single direction in the residual stream of the model.[27]

---

[26] Jailbreak refers to the process of removing restrictions set on a specific system. In this case, an LLM assistant.

[27] 🔗 alignmentforum.org/posts/jGuXSZgv6qfdhMCuJ/refusal-in-llms-is-mediated-by-a-single-direction

Something that reinforces the idea that inference-time guardrails remain some of the only (pseudo) effective ways to limit the behavior of systems.[28]

Another observation worth mentioning is that preserving other complex values may require solutions outside the framework of techniques we used so far. If we consider truthfulness and honesty, engineering a reward signal or dataset to represent "truthful information" is not (currently) possible. And given the propensity of ML models to hallucinate, the challenge becomes more complex. Even though results on benchmarks like TruthfulQA [358] and FactCheckQA [29] tend to improve as models are better trained and fine-tuned, i.e., receive more high-quality-factual data, hallucination may be an inevitable feature of LLM trained via a learning paradigm [629]. Hence, impact mitigation guardrails that protect this principle (truthfulness) probably require a non-DL solution.[29]

To summarize, the creation of guardrails presents both contextual and fundamental challenges. The uniqueness of the impact mitigation strategy will depend on the specific side effects we aim to address and the principles we adhere to. Nevertheless, we contend that neglecting this crucial step — defining external sources of alignment beyond human preferences but within our societal norms — is not an option if we seek to approach any level of robust value alignment. Meanwhile, as a concluding remark for this final section, we would like to point to some avenues

---

[28] But only in cases where the guardrail models are not themselves susceptible to adversarial exploitations.

[29] In fact, aligning ML systems with this principle may intuitively require the same kind of external structure we have. In general, a single human actor has limited misinformation detection capabilities besides its internal sense of skepticism, which can be biased in many ways. However, given the ample amount of stored and curated knowledge available and the cultural mechanism we implemented to check and assess the validity of information (fact-checking [473], flagging [200], crowdsourcing [585]), we can (trough some effort) find the grain of truth in many cases. But to implement this in our current paradigm requires us to be open to more hybrid approaches, taking us back to ideas like knowledge bases and symbolic AI [350, 236, 6, 290, 56]. However, augmenting these models with tools for information retrieval, web searching, and other functionalities like code execution, besides solving specific problems, creates new challenges we do not face when working with unaugmented language models [400].

of future research that could help improve future impact mitigation strategies:

- Few resources represent high-quality red teaming efforts, especially in non-English languages. The available datasets of this study have a severe sample imbalance concerning language, making our Brazilian Portuguese (auxiliary) preference model an unreliable proxy. **Crating more high-quality open-source datasets can help improve future safety works in low-resource languages.**
- If scaling laws apply to preference modeling, **training bigger open-source preference models should help future alignment works.**
- The same can be said for evaluation benchmarks. Currently, most benchmarks focused on safety are predominantly in English, hindering the contextual understanding of vulnerabilities ingrained in non-English Languages. **Creating non-English evaluation harnesses tailored for alignment can also help increase advances in research (as general interest towards alignment) on the topic worldwide.**
- Addressing biases in preference modeling is also an avenue of research. For example, our models seem to "forgive" toxic samples if the affected target group represents a non-minority. Thus, if we wish to prevent these models from producing toxicity independent of the targeted group, **more diverse datasets and red teaming are required.**
- It is unclear if a very general preference model or several specific preference models, aggregated with a fine-tuned selection of weights, is the best approach to alignment. **Investigating how to aggregate several learning signals in a learning dynamic capable of preserving commonalities and significant differences is a valuable avenue of research for alignment theorists seeking to tackle the problem of preference aggregation.**

- Much of what is done in AI safety and alignment is not standardized. **Creating an overarching framework for impact mitigation can help us standardize what ML guardrails should look like and what minimum requirements they should satisfy.**
- What kinds of problems are out of reach from DL-based solutions? **Identifying what issues require a specific mode of work that would fall outside the ML paradigm could help us better understand the limitations of the learning paradigm in terms of safety and alignment.**
- Vulnerability to adversarial attacks is an Achilles heel of all DL-based systems. **Learning how to create guardrails that can detect adversarial attacks or out-of-distribution scenarios seems necessary,** given the prospect of aligned models is currently one jailbreak away from being a do-anything-now-system.
- Rejection sampling is one of the most cost-effective, low-fiddling strategies for creating guardrails. With advents like LangChain, there is much room for exploration on **creating reflexive and chain-of-thought style guardrails that can reason on a more sophisticated level about what should and should not get outside the box.**

7.5 Epilogue

Preventing general-purpose ML systems from behaving in an unwanted way is a difficult task. Given that one of the premises of this paradigm is that we do not specify every imperative act that should be done, much of their behavior remains a mystery till it emerges. And when tests do not exhaust the search space, unwanted behaviors are only spotted in the wild, compromising the safety of those involved. Thus, given that we aim to develop general-purpose AI systems and generality entails systems with a vast output space, defining impact mitigation methods should become a part of any developmental cycle for systems that require alignment.

Given that aligning AI systems entails not only the blind following of any human desire but also the prevention of harm,[30] alignment theorists and engineers must be mindful of this stage. And if we consider the suppositions behind Dynamic Normativity, we can subscribe to the ideas that part of our normativity lives outside. This means we can work in alignment from two different ends. One focused on the preferences of individuals, and the other on the restrictions imposed by society. Ultimately, the helpful versus harmless tension can be thought of as an expression of the all-to-human experience of "*doing what I want versus doing what I am allowed*." Something that we also struggle to deal with in many situations of our lives.

To impose such restrictions, we reviewed a series of impact mitigation measures. These measures can be implemented before training (creating high-quality datasets for pretraining), during training (adding penalty terms in the optimized objective function), and during inference (adding guardrails for the deployed model). To present a minimal implementation of what impact mitigation tools for general-purpose ML-based language assistants would look like, we developed two auxiliary preference models to serve as complementary sources in a learning/impact mitigation dynamic. As a side note, at the moment of this writing, all models and datasets tied to this project collectively possess more than 200,000 downloads, which we take as small proof that this kind of open work, with the intent to open and democratize research in alignment, is welcomed by the community, giving us more reasons to push this agenda of research forward.

Nevertheless, numerous problems and challenges persist, offering several paths for future investigation. While our implementation of Dynamic Normativity falls

---

[30] Harms, we at least agree, in a democratic sense, on the harmfulness.

short of resolving value alignment, we argue that our theory points us to the necessary conditions required so that a learning approach can conquer this predicament. Moreover, it delineates the additional prerequisites necessary to approximate this objective. While necessary conditions are, one could say, philosophical foundations, sufficient conditions are a three-stage engineering blueprint. Each stage has its specific challenges, and the more we can conquer them, the closer we get to value alignment.

Finally, as we move closer to value alignment, we also move closer to developing systems that can safely be integrated into human society, serving as extensions of our own will and potentially, in the future, evolving to the status of *"our allies"* and not just tools. In addition to placing human volition as the guiding compass of these systems, alignment is also about discovering ways in which we and our creations can better cooperate and promote human flourishing. We hope this work can help us bring this goal closer to fruition.

Closing Remarks

"A conclusion is simply the place where you got tired of thinking".

— Dan Chaon, Stay Awake

At the beginning of this work, we proposed to address the question "What should AI systems do?" from both a philosophical and an engineering perspective. Besides aiming to be interdisciplinary researchers, i.e., those who build bridges between different areas of knowledge, we argue that this normative question demands this type of endeavor. Techno-humanistic. Techno-philosophical. A perspective that relates to the dawn of AI research. A time when the development of AI systems was not merely a technical endeavor but an inherently philosophical one, demanding consideration of deep questions related to the nature of cognition.

In our opinion, posing the normative question allows for a comeback of this techno-humanistic approach, given that the study of normativity remains, to some extent, a major philosophical project. Trying to address the normative challenges that emerge from this advancing field without the aid of philosophy would be to ignore the advantage that "giant shoulders" can give us. Meanwhile, without the technical background and implementational output, we fall victim to much of what we have criticized at the beginning of this work: hollow, abstract, and ungrounded ethical discourse unable to impose its normativity in praxis.

This claim came as an interpretation of the descriptive work of Worldwide AI Ethics,[31] where instead of solely relying on the work of others to paint a picture of "What AI Ethics is about?" we sought to create an original analysis to improve our understanding of the normative discourse of this interdisciplinary field. This analysis sought to surpass past descriptive works in many points, like sample size, richness in investigation, data visualization, and the openness of how we presented our results to the community. At the time of this writing, our work remains the most extensive meta-analytical investigation with AI ethics guidelines, which undeniably have helped shape us to shape our understanding of the resonances and conflicts of this field. Something we hope the readers can also come to experience.

For the resonances, we could cite the revelation that certain principles dominate the normative discourse in this community. For conflicts, we can mention the diversity in which these principles can be described and defined. These and other conclusions were instrumental in driving the progress of this research since they showed us points of fragility we could try to contribute, as well as the inherent paradoxes AI Ethics, and perhaps all interdisciplinary fields, face. Then, after spending the first portion of this work looking for "what values should guide the use and development of AI systems?", we focused the rest of this work on the problem of designing AI systems that incorporate such values from the outset by developing a framework capable of fostering the creation of a specialized form of ethical AI: *aligned AI*.

Alignment is an intriguing problem. Given the specificity of learning-based techniques, the foundation of our deep learning paradigm, instead of precisely coding a solution, we create a signal that can help an optimizer find the best-suiting model for our problem via iterated (gradient-based) updates. Given the diffi-

---

[31] 🔗 nkluge-correa.github.io/worldwide_AI-ethics/

culties related to expressing precise objectives and all vulnerabilities associated with ML (opaqueness, brittleness, etc.), many believe that if we create AGI via this paradigm, we might come to lose control over our systems. From this belief comes the idea that we need to find ways to guarantee the controllability of this paradigm's by-products, especially those of considerable capability and generality.

While some relate the control problem to the inherent characteristics of intelligent behavior and instrumentality, we propose that alignment does not have to rely on these assumptions or the very uncertain timelines regarding AGI and X-risk scenarios to be considered a problem worth investigating. Even if speculations about future technologies are exciting possibilities to let one's mind wander, alignment is not just about future AI but about the current systems that we are starting to let into our lives. Thus, in this work, we sought to define the alignment problem as, besides being a desirable solution to the control problem, a symptom of the learning paradigm applied to neural networks. A symptom related to the challenges of guaranteeing that a neural network-based model found by a gradient-based optimizer will carry the objectives and intentions specified by the controllers. Meanwhile, as many examples have shown us, many situations in which models operate poorly in the wild are alignment failures, i.e., objective functions that do not embed all relevant aspects of the environment we value generate systems that optimize for something we did not want.

These implications have more severe consequences when the output of these models has moral implications. Also, as these systems become more general and capable, the side effects of unwanted behaviors may become more pronounced. For example, misinformation at an individual level is nothing compared to deceiving an entire society about a vital fact. Hence, with safety and prudence in mind, we can start to work on these issues with our modestly general and limited systems. Something that sounds easy but poses a real challenge to any ML engineer or alignment researcher.

A less noble motivation for this study is the blunt intellectual trill open problems can cause. Alignment offers a convergence point for many areas of research. Just like the beginning of AI research was a meeting ground for computer scientists, mathematicians, cognitive scientists, and many other intellectuals, alignment offers the same. An interdisciplinary playground where many can come and contribute.

As a philosophical endeavor, we sought to support our proposal with a collection of metaphysical and metaethical underpinnings, serving as the roots of our work. Every theory or hypothesis has its foundations, and while some authors choose to conceal or merely suggest them as implicitly stated, we aim for the opposite. Consequently, we explicitly articulated all foundational perspectives on how this work conceptualizes intelligence and normativity from a dynamic standpoint and how we constructed a framework called Dynamic Normativity. This framework attempts to reconstruct, in a limited way, the human normative experience, which is, to a great extent, a learning experience.

In Dynamic Normativity, we hypothesize that a specific set of necessary conditions must hold to address value alignment. Even though insufficient to ensure said alignment, we advocated for their essentiality as irrevocable facts that should be for a learning-based approach toward alignment to succeed.

1. Goals are fundamental aspects of intelligent and intentional behavior.
2. Intentions permeate human behavior.
3. Normative preferences permeate human intentions.
4. Through actions, humans impregnate their environment with the preferences they possess.

From these, we postulate that a system that accurately models human intentions and the human environment can indirectly access normative information embedded in them. These are speculations on the capabilities of the learning paradigm and the nature of human intentionality and normative behavior. Following this claim, we proposed a set of sufficient conditions to help us establish a minimum set of criteria for value alignment:

1. Aligned AI systems should coherently aggregate human preferences in a way that resolves cases of uncertainty. Aligning AI systems requires methods to deal with cases of uncertainty.
2. AI systems can adhere to human preferences if they are an available part of their objective function. Aligning AI systems requires using human preferences as part of their learning signal.
3. Aligned AI systems should have mechanisms to perform impact mitigation to minimize harmful and unintended consequences. Aligning AI systems requires the specification of safety guardrails.

We argue that no value alignment can come to fruition if the necessary conditions do not hold. Meanwhile, how well a system is aligned depends on how well the sufficient conditions are satisfied. These conditions give us a starting point to build aligned systems from an applied and theoretical perspective. All our developments (the Aira series)[32] are inspired by this proposed framework.

However, numerous problems persist, and each step of this process is fraught with difficulties. Collecting human data and feedback is a controversial process that follows unethical practices in many circumstances. Aggregating human preferences possess many impossibilities and trade-offs that are still vastly unexplored from an alignment perspective. Value learning, even from high-quality sources

---

[32] 🔗 huggingface.co/collections/nicholasKluge/aira-657db1563c65a5be2a02f51c

of demonstrations and feedback, still produces unwanted behavior and emergent properties that are hard to foresee. Known impact mitigation techniques do not provide a full prof methodology to prevent these undesirable behaviors. Finally, to make everything worse, much of the state-of-the-art work on these problems comes in a way that is not reproducible to much of the community due to (1) resource costs and (2) the closed-way current research is being done.

While lowering the entrance bar in terms of resources and knowledge for specialized fields might not be something we can improve too much on, there is (definitely) something to be said about making our current scenario (especially in Academia) more open in terms of research, development, and sharing of resources. Something that brings the perils of unalignment back to the human sphere.

One of the most significant challenges of alignment is dealing with human alignment. We did not address this factor in this work, but it represents a substantial portion of the problem. Nowadays, and till AI systems gain substantially more autonomy, most of the dangers related to these technologies come from people utilizing them in an undesirable way. Making AI ethics, AI safety, and AI alignment literacy and research obscure or guarded behind doors can not help improve this scenario. At the inception of this work in 2020, there was a dearth of resources and limited general interest in these research areas. However, presently, the community is gradually establishing a field. This work represents a modest interdisciplinary contribution to bridging the knowledge and resource gap in areas where these topics are still inadequately comprehended. Something that remains to the reader to judge if we have been successful or not.

As a famous thinker once said, the most important of all human problems is the "*moral problem*". If this is the most crucial problem, would it be overly audacious to state that the alignment problem is the paramount issue involving AI? If the answer is affirmative, we could say that this is one of the "goods" alignment has brought: turning AI research into a more humane science and a common place to think about interesting things and work together. Something we hope this work can represent.

Acknowledgements

First and foremost, I would like to extend my most profound appreciation to my family for their endless support and encouragement. A special dedication goes to my grandmother, **Maria Tereza Dutra Kluge**, who passed away in May 2024.

I also want to thank my advisors, Nythamar de Oliveira and Michael Schulz, for their unwavering support, guidance, and encouragement throughout my doctoral journey. I am also profoundly grateful to the institutions that made this research possible: the University of Bonn and the Pontifical Catholic University of Rio Grande do Sul. I would also like to thank the Coordenação de Aperfeiçoamento de Pessoal Nível Superior – Brazil (CAPES), the Deutscher Akademischer Austauschdienst (DAAD), and the Rede de Inteligência Artificial Ética e Segura (RAIES) for their financial support.

My heartfelt thanks go to my AI Robotics Ethics Society (AIRES) colleagues, who have collaborated with me these last few years. Your collaboration, insights, and camaraderie have been fundamental to realizing many projects I am glad to have been involved in.

Lastly, I want to acknowledge all the wonderful, albeit slightly crazy, individuals who cohabitate with me at the CPP, close to Bonn HBF.

I love you all.

Bibliography

[1] Scott Aaronson. Why philosophers should care about computational complexity. *Computability: Turing, Gödel, Church, and Beyond*, 261:327, 2013.

[2] Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*, 2023.

[3] Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546, 1953.

[4] James Allen, Nate Blaylock, and George Ferguson. A problem solving model for collaborative agents. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 774–781, 2002.

[5] Gordon Willard Allport. Personality: A psychological interpretation. 1937.

[6] Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*, pages 468–485. PMLR, 2022.

[7] William P Alston. *Epistemic justification: Essays in the theory of knowledge*. Cornell University Press, 1989.

[8] Joan Alway and Craig Calhoun. Critical social theory: Culture, history, and the challenge of difference. 26(1):119, 1 1997.

[9] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. 06 2016.

[10] Philip W Anderson. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972.

[11] Darko Andročec. Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, 12(2):205–216, 2020.

[12] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947, 2024.

[13] Stuart Armstrong. Chaining god: A qualitative approach to ai, trust and moral systems. *Unpublished manuscript, October*, 20:2010–1, 2007.

[14] Stuart Armstrong, Anders Sandberg, and Nick Bostrom. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines*, 22(4):299–324, 2012.

[15] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

[16] Elizabeth Ashford and Tim Mulgan. Contractualism. 2007.

[17] Isaac Asimov. *I, robot*, volume 1. Spectra, 2004.

[18] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom

Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[19] David Auerbach. The most terrifying thought experiment of all time. *Slate Magazine*, 2014.

[20] Caglar Aytekin. Neural networks are decision trees. *arXiv preprint arXiv:2210.05189*, 2022.

[21] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[22] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[23] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[24] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

[25] Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*, 2021.

[26] Ed Barbeau. Fallacies, flaws, and flimflam. *College Mathematics Journal*, pages 149–154, 1993.

[27] Anthony M Barrett and Seth D Baum. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):397–414, 2017.

[28] Andrew Basden. On the application of expert systems. *International Journal of Man-Machine Studies*, 19(5):461–477, 1983.

[29] Vasilisa Bashlovkina, Zhaobin Kuang, Riley Matthews, Edward Clifford, Yennie Jun, William W Cohen, and Simon Baumgartner. Trusted source alignment in large language models. *arXiv preprint arXiv:2311.06697*, 2023.

[30] Jean Baudrillard. *Simulacra and Simulation*. 1994.

[31] Jean Baudrillard. *The Consumer Society: Myths and Structures*. 1998.

[32] Seth Baum. A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, pages 17–1, 2017.

[33] Seth D Baum. Superintelligence skepticism as a political tool. *Information*, 9(9):209, 2018.

[34] Seth D Baum, Ben Goertzel, and Ted G Goertzel. How long until human-level ai? results from an expert assessment. *Technological Forecasting and Social Change*, 78(1):185–195, 2011.

[35] CWJ Beenakker. Hempel's dilemma and the physics of computation. *arXiv preprint physics/0702072*, 2007.

[36] Randall D Beer. Dynamical approaches to cognitive science. *Trends in cognitive sciences*, 4(3):91–99, 2000.

[37] Randall D Beer. Dynamical systems and embedded cognition. *The Cambridge handbook of artificial intelligence*, (812):856–873, 2014.

[38] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[39] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[40] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[41] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[42] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023.

[43] Mounir Bensalah. Toward an ethical code of ai and human rights in morocco. 1, 2021.

[44] Anthony Berglas. Artificial intelligence will kill our grandchildren (singularity). *Unpublished manuscript, draft*, 9, 2012.

[45] Paul Bernays. Alonzo church. an unsolvable problem of elementary number theory. american journal of mathematics, vol. 58 (1936), pp. 345–363. *The Journal of Symbolic Logic*, 1(2):73–74, 1936.

[46] Ram Bhadra. Linkedin: A case study into how tech giants like microsoft abuse their dominant market position to create unlawful monopolies in emerging industries. *Hastings Sci. & Tech. LJ*, 13:3, 2022.

[47] Alfred Binet and Theophile Simon. The development of intelligence in children. 1961.

[48] Duncan Black. On the rationale of group decision-making. *Journal of political economy*, 56(1):23–34, 1948.

[49] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, et al. Interpreting neural networks through the polytope lens. *arXiv preprint arXiv:2211.12312*, 2022.

[50] Simon Blackburn. *Ruling passions*. Oxford University Press Oxford, 1998.

[51] N Block. ,'psychologism and behaviorism', philosophical review, 90: 5-43. 1981.

[52] Allan Bloom, Adam Kirsch, et al. *The republic of Plato*, volume 2. basic books New York, 1968.

[53] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[54] Mohammad Reza Bonyadi and Zbigniew Michalewicz. Particle swarm optimization for single objective continuous space problems: a review. *Evolutionary computation*, 25(1):1–54, 2017.

[55] S Borgeaud, A Mensch, J Hoffmann, T Cai, E Rutherford, K Millican, G Driessche, JB Lespiau, B Damoc, A Clark, et al. Improving language models by retrieving from trillions of tokens. arxiv. *arXiv preprint arXiv:2112.04426*, 2021.

[56] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine*

learning, pages 2206–2240. PMLR, 2022.

[57] Leonie N Bossert and Thilo Hagendorff. The ethics of sustainable ai: Why animals (should) matter for a sustainable use of ai. *Sustainable Development*, 2023.

[58] Nick Bostrom. Existential risks: Analyzing human extinction scenarios and related hazards. 15, 2002.

[59] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 2014.

[60] Ben Bradley. Doing away with harm. *Philosophy and Phenomenological Research*, 85(2):390–412, 2012.

[61] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[62] Arthur P Brief. Values and codes of conduct on fraudulent financial reporting. *Journal of Business Ethics*, 15(2):183–198, 1996.

[63] Markus Brill and Felix Fischer. The price of neutrality for the ranked pairs method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1299–1305, 2012.

[64] Gerald W Brock. *Telecommunication policy for the information age: From monopoly to competition*. Harvard University Press, 1994.

[65] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[66] John Brunner. *Stand on Zanzibar*. 1968.

[67] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke,

Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[68] Bruce G Buchanan and Edward H Shortliffe. *Rule-based expert systems: the MYCIN experiments of the Stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[69] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[70] Claude J Burtenshaw. The political theory of pluralist democracy. *Western Political Quarterly*, 21(4):577–587, 1968.

[71] Samuel Butler. Darwin among the machines, 1863.

[72] Krister Bykvist. Moral uncertainty. *Philosophy Compass*, 12(3):e12408, 2017.

[73] Terrell Ward Bynum. Flourishing ethics. *Ethics and information technology*, 8(4):157–173, 2006.

[74] Ryan Calo. Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, 51:399, 2017.

[75] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[76] Víctor Campos, Pablo Sprechmann, Steven Hansen, Andre Barreto, Steven Kapturowski, Alex Vitvitskyi, Adria Puigdomenech Badia, and Charles Blundell. Beyond fine-tuning: Transferring behavior in reinforcement learning. *arXiv preprint arXiv:2102.13515*, 2021.

[77] Ryan Carey. Incorrigibility in the cirl framework. In *Proceedings of the 2018*

AAAI/ACM Conference on AI, Ethics, and Society, pages 30–35, 2018.

[78] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. *arXiv preprint arXiv:2403.06634*, 2024.

[79] John B Carroll et al. *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge University Press, 1993.

[80] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[81] Davide Castelvecchi et al. Hunt for gravitational waves to resume after massive upgrade. *Nature*, 525(7569):301–302, 2015.

[82] Raymond B Cattell and Paul Ed Kline. *The scientific analysis of personality and motivation.* Academic Press, 1977.

[83] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[84] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[85] David Chalmers. The singularity: A philosophical analysis. 17, 2010.

[86] David J Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks, 1996.

[87] David J Chalmers. *The conscious mind: In search of a fundamental theory*. Ox-

ford Paperbacks, 1997.

[88] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[89] Frances Chapman and Lauren Tarasuk. Slender man on trial: Has media taken the minds of the young? *Criminal Law Bulletin*, 52(3), 2016.

[90] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[91] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[92] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

[93] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[94] Jaedeug Choi and Kee-Eung Kim. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. *Advances in neural information processing systems*, 25, 2012.

[95] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[96] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[97] Brian Christian. *The alignment problem: How can machines learn human values?* Atlantic Books, 2021.

[98] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[99] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[100] Patricia Smith Churchland. *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press, 1989.

[101] Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, et al. Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*, 2024.

[102] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[103] Margaret Anne Cleek and Sherry Lynn Leonard. Can corporate codes of ethics influence behavior? *Journal of business ethics*, 17(6):619–630, 1998.

[104] Michael Cohen, Badri Vellambi, and Marcus Hutter. Asymptotically unambitious artificial general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2467–2476, 2020.

[105] Paul R Cohen and Adele E Howe. How evaluation guides ai research: The

message still counts more than the medium. *AI magazine*, 9(4):35–35, 1988.

[106] Patrick Collison and Michael Nielsen. Science is getting less bang for its buck. *The Atlantic*, (16), 2018.

[107] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[108] Joseph Corabi. Superintelligent ai and skepticism. *Journal of Ethics and Emerging Technologies*, 27(1):4–23, 2017.

[109] Nicholas Corrêa and Nythamar Fernandes De Oliveira. Counterfactual analysis by algorithmic complexity: A metric between possible worlds. *Manuscrito*, 45:1–35, 2022.

[110] Nicholas Kluge Corrêa. A hipótese da permissibilidade. 2021.

[111] Nicholas Kluge Corrêa and Nythamar De Oliveira. Singularity and coordination problems: Pandemic lessons from 2020. *arXiv preprint arXiv:2010.07018*, 2020.

[112] Nicholas Kluge Corrêa, Nythamar De Oliveira, and Diogo Massmann. On the efficiency of ethics as a governing tool for artificial intelligence. *arXiv preprint arXiv:2210.15289*, 2022.

[113] Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar De Oliveira. Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *Machine Learning With Applications*, 2024.

[114] Nicholas Kluge Corrêa, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, Edmund Terem, and Nythamar Oliveira. Worldwide ai ethics: a review of 200 guidelines and recommendations for ai gov-

ernance. *Patterns*, 4(10), October 2023.

[115] Nicholas Kluge Corrêa, James William Santos, Camila Galvão, Marcelo Pasetti, Dieine Schiavon, Faizah Naqvi, Robayet Hossain, and Nythamar De Oliveira. Crossing the principle–practice gap in ai ethics with ethical problem-solving. *AI and Ethics*, pages 1–18, 2024.

[116] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. Mula: a sparse mixture of experts language model trained in brazilian portuguese. `https://huggingface.co/MulaBR`, 2024.

[117] Nicholas Corrêa and Nythamar Fernandes de Oliveira. Good AI for the present of humanity democratizing AI governance. 2(2), 7 2021.

[118] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, pages 1–12, 2023.

[119] Robin Cowan. Expert systems: aspects of and limitations to the codifiability of knowledge. *Research Policy*, 30(9):1355–1372, 2001.

[120] Joseph Cox. Gpt-4 hired unwitting taskrabbit worker by pretending to be 'vision-impaired' human. *Vice: Motherboard*, 2023.

[121] Andrew Critch and David Krueger. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*, 2020.

[122] Penny Crofts and Honni van Rijswijk. Negotiating'evil': Google, project maven and the corporate form. *Law, Tech. & Hum.*, 2:75, 2020.

[123] Maria da Conceição Tavares. Economia política | maria da conceição tavares | 08, Feb. 2017.

[124] Robert A Dahl. *Dilemmas of pluralist democracy: Autonomy vs. control*. Yale University Press, 1983.

[125] Wei Dai. Towards a new decision theory. *Less Wrong*, 2009.

[126] Parag Pravin Dakle, SaiKrishna Rallabandi, and Preethi Raghavan. Understanding bloom: An empirical study on diverse nlp tasks. *arXiv preprint arXiv:2211.14865*, 2022.

[127] Norman Daniels. Wide reflective equilibrium and theory acceptance in ethics. *The journal of philosophy*, 76(5):256–282, 1979.

[128] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[129] Alison Darcy, Aaron Beaudette, Emil Chiauzzi, Jade Daniels, Kim Goodwin, Timothy Y Mariano, Paul Wicks, and Athena Robinson. Anatomy of a woebot®(wb001): agent guided cbt for women with postpartum depression. *Expert Review of Medical Devices*, 19(4):287–301, 2022.

[130] Richard B Darlington. Minimax is the best electoral system after all. *arXiv preprint arXiv:1606.04371*, 2016.

[131] Herbert Aron David. *The method of paired comparisons*, volume 12. London, 1963.

[132] Janet Davis and Lisa P Nathan. Value sensitive design: Applications, adaptations, and critiques. *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*, pages 11–40, 2015.

[133] Mike Davis. *Beyond Blade Runner: Urban Control, The Ecology of Fear*. 1992.

[134] Richard Dawkins et al. *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company, 1996.

[135] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint*

arXiv:1809.04444, 2018.

[136] Liesbeth De Mol. Turing machines. 2018.

[137] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

[138] Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2019.

[139] Daniel Dennett. Intentional systems theory. 2009.

[140] Daniel C Dennett. True believers: The intentional strategy and why it works. 1981.

[141] Daniel C Dennett. *Consciousness explained*. Penguin uk, 1993.

[142] Daniel C Dennett. Darwin's dangerous idea. *The Sciences*, 35(3):34–40, 1995.

[143] Daniel C Dennett. *Kinds of minds: Toward an understanding of consciousness*. Basic Books, 2008.

[144] Daniel Clement Dennett. *The intentional stance*. MIT press, 1987.

[145] Daniel Clement Dennett. *Sweet dreams: Philosophical obstacles to a science of consciousness*. MIT press, 2005.

[146] René Descartes. *Discourse on the method: And, meditations on first philosophy*. Yale University Press, 1996.

[147] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

[148] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert:

Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[149] John Dewey. *Experience and nature*, volume 471. Courier Corporation, 1958.

[150] Philip K. Dick. *Do Androids Dream of Electric Sheep?* 1968.

[151] Virginia Dignum. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1):1–3, March 2018.

[152] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

[153] Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

[154] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[155] Fred I Dretske. The intentionality of cognitive states. *Midwest Studies in Philosophy*, 5:281–294, 1980.

[156] K Eric Drexler. *Engines of creation*. Anchor books, 1986.

[157] Hubert L Dreyfus. What computers can't do: The limits of artificial intelligence. 1972.

[158] Hubert L Dreyfus. Husserl, intentionality, and cognitive science. 1982.

[159] Hubert L Dreyfus. *What computers still can't do: A critique of artificial reason*. MIT press, 1992.

[160] Hubert L Dreyfus and Stuart E Dreyfus. What artificial experts can and cannot do. *AI & society*, 6(1):18–26, 1992.

[161] John Durkin. Expert systems: a view of the field. *IEEE Intelligent Systems*, 11(02):56–63, 1996.

[162] Adrien Ecoffet and Joel Lehman. Reinforcement learning under moral uncertainty. In *International Conference on Machine Learning*, pages 2926–2936. PMLR, 2021.

[163] Nicholas Economides. Telecommunications regulation: an introduction. 2004.

[164] N Editorials. Tools such as chatgpt threaten transparent science; here are our ground rules for their use. *Nature*, 613(612):10–1038, 2023.

[165] Arpad E Elo. The proposed uscf rating system. its development, theory, and applications. *Chess Life*, 22(8):242–247, 1967.

[166] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

[167] Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*, 2020.

[168] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[169] Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. Stanford human preferences dataset, 2023.

[170] Amitai Etzioni and Oren Etzioni. Ai assisted ethics. *Ethics and Inf. Technol.*, 18(2):149–156, jun 2016.

[171] Owain Evans, Andreas Stuhlmüller, and Noah Goodman. Learning the preferences of ignorant, inconsistent agents. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[172] Tom Everitt, Gary Lea, and Marcus Hutter. Agi safety literature review. *arXiv preprint arXiv:1805.01109*, 2018.

[173] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics, 2019.

[174] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021.

[175] Andrew Feenberg. *Questioning technology*. Routledge, 2012.

[176] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[177] Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024.

[178] Matt Field. In syria, russia found the chance to showcase its swagger–and its robot weapons. *Bulletin of the Atomic Scientists*, 2019.

[179] Roderick Firth. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, 12(3):317–345, 1952.

[180] McKenna Fitzgerald, Aaron Boddy, and Seth D Baum. 2020 survey of artificial general intelligence projects for ethics, risk, and policy. 2020.

[181] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. 2020.

[182] Jerry A Fodor. The mind-body problem. *Scientific american*, 244(1):114–123, 1981.

[183] Philippa Foot. Does moral subjectivism rest on a mistake? *Royal Institute of Philosophy Supplements*, 46:107–123, 2000.

[184] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*, 2020.

[185] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104, 2019.

[186] Henry Fountain. A dream of clean energy at a very high price. *New York Times*, 3:17, 2017.

[187] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[188] Edward Fredkin. An introduction to digital philosophy. *International journal of theoretical physics*, 42(2):189–247, 2003.

[189] Carl Benedikt Frey and Michael Osborne. The future of employment. 2013.

[190] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[191] Thomas Fuchs. *Ecology of the brain: The phenomenology and biology of the embodied mind*. Oxford University Press, 2017.

[192] Francis Fukuyama, Barak Richman, and Ashish Goel. How to save democracy from technology: ending big tech's information monopoly. *Foreign Aff.*,

100:98, 2021.

[193] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[194] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.

[195] Shaun Gallagher. *How the body shapes the mind*. Clarendon Press, 2006.

[196] Shaun Gallagher and Matthew Bower. Making enactivism even more embodied. 2013.

[197] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

[198] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[199] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.

[200] Dongfang Gaozhao. Flagging fake news on social media: An experimental study of media consumers' identification of fake news. *Government Informa-*

tion Quarterly, 38(3):101591, 2021.

[201] Eduardo A. S. Garcia. Open portuguese llm leaderboard. `https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard`, 2024.

[202] Eugenio V Garcia. The militarization of artificial intelligence: a wake-up call for the global south. *Available at SSRN 3452323*, 2019.

[203] H Gardner. Frames of mind theory of multiple intelligences. developing talent in young people, 1983.

[204] Scott Garrabrant, Daniel A Herrmann, and Josiah Lopez-Wild. Cartesian frames. *arXiv preprint arXiv:2109.10996*, 2021.

[205] David Gauthier. Why contractarianism? *ETHICA*, page 571, 2013.

[206] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[207] Edward Moore Geist. It's already too late to stop the ai arms race—we must manage it instead. *Bulletin of the Atomic Scientists*, 72(5):318–321, 2016.

[208] Allan Gibbard. *Wise choices, apt feelings: A theory of normative judgment*. Harvard University Press, 1990.

[209] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[210] William Gibson. *Neuromancer*. Sprawl trilogy. 1984.

[211] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[212] Gary SW Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and

Alexander Binder. Understanding integrated gradients with smoothtaylor for deep neural network attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4949–4956. IEEE, 2021.

[213] Irving John Good. On the marking of chess-players. *The Mathematical Gazette*, 39(330):292–296, 1955.

[214] Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.

[215] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[216] Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pages 91–121. Springer, 1984.

[217] Linda S Gottfredson. g: Highly general and highly practical. In *The general factor of intelligence*, pages 343–392. Psychology Press, 2002.

[218] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

[219] Edward J Gracely. On the noncomparability of judgments made by different ethical theories. *Metaphilosophy*, 27(3):327–332, 1996.

[220] David Graeber. *Direct action: An ethnography*. AK press, 2009.

[221] David Graeber. *The democracy project: A history, a crisis, a movement*. Random House, 2013.

[222] Harvey J Graff. *Undisciplining knowledge: Interdisciplinarity in the twentieth century*. JHU Press, 2015.

[223] Harvey J Graff. The “problem” of interdisciplinarity in theory, practice, and history. *Social Science History*, 40(4):775–803, 2016.

[224] Monica Greco. On the vitality of vitalism. *Theory, Culture & Society*, 22(1):15–27, 2005.

[225] James Green-Armytage. Four condorcet-hare hybrid methods for single-winner elections. *Voting matters*, 29:1–14, 2011.

[226] Richard L Gregory and Oliver Louis Zangwill. *The Oxford companion to the mind.* Oxford university press, 1987.

[227] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*, 2023.

[228] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023.

[229] Tricia A Griffin, Brian P Green, and Jos VM Welie. The ethical wisdom of ai developers. *AI and Ethics*, pages 1–11, 2024.

[230] Ross Gruetzemacher, David Paradice, and Kang Bok Lee. Forecasting extreme labor displacement: A survey of ai practitioners. *Technological Forecasting and Social Change*, 161:120323, 2020.

[231] Michael M. Grynbaum and Ryan Mac. The times sues openai and microsoft over a.i. use of copyrighted work. *The New York Times*.

[232] Joy Paul Guilford. The nature of human intelligence. 1967.

[233] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[234] Bin Guo, Hao Wang, Yasan Ding, Wei Wu, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. Conditional text generation for harmonious human-machine interaction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–50, 2021.

[235] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

[236] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[237] Arthur Gwagwa. Reccomendations on the inclusion subsaharan africa in global ai ethics, 2019.

[238] Arthur Gwagwa, Patti Kachidza, Kathleen Siminyu, and Matthew Smith. Responsible artificial intelligence in sub-saharan africa : landscape and general state of play, 2021.

[239] Susan Haack. *Evidence and inquiry: towards reconstruction in epistemology*. Blackwell, Oxford, UK; Cambridge, USA, 1993.

[240] Susan Haack and William Duica. Evidence and inquiry. towards reconstruction in epistemology. *Ideas y Valores*, 46(104):99–105, 1997.

[241] Michael Carl Haas and Sophie-Charlotte Fischer. The evolution of targeted killing practices: Autonomous weapons, future conflict, and the international order. *Contemporary Security Policy*, 38(2):281–306, 2017.

[242] Selcan Hacaoglu. Turkey's killer drone swarm poses syria air challenge to putin. *Bloomberg News*, 2020.

[243] Steven C Hackett. Incomplete contracting: a laboratory experimental anal-

ysis. *Economic Inquiry*, 31(2):274–297, 1993.

[244] Dylan Hadfield-Menell and Gillian K Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

[245] Thilo Hagendorff. The ethics of AI ethics – an evaluation of guidelines. 02 2019.

[246] Thilo Hagendorff, Leonie N Bossert, Yip Fai Tse, and Peter Singer. Speciesist bias in ai: how ai applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics*, pages 1–18, 2022.

[247] D Hambling. Turkish military to receive 500 swarming kamikaze drones, 2020.

[248] Garrett Hardin. Extensions of" the tragedy of the commons". *Science*, 280(5364):682–683, 1998.

[249] Mark Harris. Amazon's mechanical turk workers protest:'i am a human being, not an algorithm'. *The Guardian*, 3, 2014.

[250] John C Harsanyi. Morality and the theory of rational behavior. *Social research*, pages 623–656, 1977.

[251] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

[252] John Haugeland. *Artificial intelligence: The very idea*. MIT press, 1989.

[253] Martin Heidegger. *Being and time*. Suny Press, 2010.

[254] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive prefence learning: Learn-

ing from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

[255] Carl Hempel. Reduction: Ontological and linguistic facets. 1969.

[256] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

[257] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. 09 2021.

[258] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

[259] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[260] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48:397–447, 2017.

[261] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[262] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

[263] Christopher S Hill. Imaginability, conceivability, possibility and the mind-body problem. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 87(1):61–85, 1997.

[264] Thomas Hobbes. *Hobbes's leviathan*. 1967.

[265] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[266] Jonathan K Hodge and Richard E Klima. *The mathematics of voting and elections: a hands-on approach*, volume 30. American Mathematical Soc., 2018.

[267] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[268] Ted Honderich. *The Oxford companion to philosophy*. OUP Oxford, 2005.

[269] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

[270] Robert Hooke and Terry A Jeeves. "direct search" solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229, 1961.

[271] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[272] Eric Horvitz. One hundred year study on artificial intelligence: reflections and framing. *Microsoft. com*, 2014.

[273] Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. *arXiv preprint arXiv:1606.06356*, 2016.

[274] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[275] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[276] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[277] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. 06 2019.

[278] James L Hudson. Subjectivization in ethics. *American Philosophical Quarterly*, 26(3):221–229, 1989.

[279] Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

[280] David Hume. *A treatise of human nature*. Courier Corporation, 2003.

[281] David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

[282] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[283] Edmund Husserl. *Ideen zu einer Phänomenologie und phänomenologischen Philosophie*. M. Niemeyer, 1922.

[284] Edmund Husserl and Dermot Moran. *Logical Investigations Volume 1*. Routledge, 2012.

[285] Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on*

algorithmic probability. Springer Science & Business Media, 2004.

[286] Daniel D Hutto and Erik Myin. *Radicalizing enactivism: Basic minds without content*. MIT press, 2012.

[287] Daniel D Hutto and Erik Myin. *Evolving enactivism: Basic minds meet content*. MIT press, 2017.

[288] Lilly Irani. The hidden faces of automation. *XRDS: Crossroads, The ACM Magazine for Students*, 23(2):34–37, 2016.

[289] STV Is. Voting matters-issue 6, may 1996.

[290] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

[291] Pierre Jacob. Intentionality. 2003.

[292] Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232, 2023.

[293] Fredric Jameson. *Postmodernism, Or, The Cultural Logic of Late Capitalism*. 1991.

[294] Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In *International Conference on Machine Learning*, pages 1645–1654. PMLR, 2017.

[295] Mohammad Ali Javidian, Pooyan Jamshidi, Marco Valtorta, and Rasoul Ramezanian. Preventing social disappointment in elections. *arXiv preprint arXiv:1801.05911*, 2018.

[296] Arthur R Jensen. The g factor: The science of mental ability. *Psycoloquy*, 10(2), 1999.

[297] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[298] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021.

[299] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*, 2022.

[300] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. Can machines learn morality? the delphi experiment. *arXiv e-prints*, pages arXiv–2110, 2021.

[301] Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. *arXiv preprint arXiv:1906.07132*, 2019.

[302] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. 1(9):389–399, 9 2019.

[303] Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*, 2023.

[304] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Auto-

matically auditing large language models via discrete optimization. *arXiv preprint arXiv:2303.04381*, 2023.

[305] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[306] Mislav Juric, Agneza Sandic, and Mario Brcic. Ai safety: state of the field through quantitative lens. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1254–1259. IEEE, 2020.

[307] Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N Asokan. A little goes a long way: Improving toxic language classification despite data scarcity. *arXiv preprint arXiv:2009.12344*, 2020.

[308] Theodore John Kaczynski. Industrial society and its future, 1995.

[309] Shelly Kagan. *Normative ethics*. Routledge, 2018.

[310] Cecilia Kang. OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*, May 2023.

[311] Elsa B Kania. Battlefield singularity. *Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS*, 2017.

[312] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[313] Christine Kaufmann. Ilo, declaration on social justice for a fair globalization. *ILO*, 2008.

[314] Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.

[315] Sountongnoma Martial Anicet Kiemde and Ahmed Dooguy Kora. Towards an ethics of AI in africa: rule of education. 2(1):35–40, 10 2021.

[316] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*, 2022.

[317] Robert Kirk. *Zombies and consciousness*. Clarendon Press, 2005.

[318] Alex Knapp. How much does it cost to find a higgs boson? *Forbes Online*, 2012.

[319] Andrei Nikolaevich Kolmogorov and Albert T Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

[320] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[321] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

[322] Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. Controlling conditional language models without catastrophic forgetting. In *International Conference on Machine Learning*, pages 11499–11528. PMLR, 2022.

[323] Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-

tuning language models with no catastrophic forgetting. *arXiv preprint arXiv:2206.00761*, 2022.

[324] Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

[325] George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*, 2023.

[326] György Kovács, Pedro Alonso, and Rajkumar Saini. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15, 2021.

[327] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

[328] Dmitrii Krasheninnikov, Rohin Shah, and Herke van Hoof. Combining reward information from multiple sources. *arXiv preprint arXiv:2103.12142*, 2021.

[329] Thomas G Krattenmaker. The telecommunications act of 1996. *Fed. Comm. LJ*, 49:1, 1996.

[330] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[331] Ray Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. 2005.

[332] Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third,

and Miriam Fernandez. Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining*, 12(1):166, 2022.

[333] Geoffrey LaForte, Patrick J Hayes, and Kenneth M Ford. Why gödel's theorem cannot refute computationalism. *Artificial Intelligence*, 104(1-2):265–286, 1998.

[334] Viet Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Huu Nguyen. Open multilingual llm evaluation leaderboard. `https://huggingface.co/spaces/uonlp/open_multilingual_llm_leaderboard`, 2023.

[335] Yann LeCun. Yann lecun: Deep learning, convnets, and self-supervised learning | lex fridman podcast #36, Aug. 2019.

[336] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[337] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

[338] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[339] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

[340] Shane Legg, Marcus Hutter, et al. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17, 2007.

[341] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg,

Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

[342] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400, 2019.

[343] Gottfried Wilhelm Leibniz. Discourse on metaphysics. In *Philosophical papers and letters*, pages 303–330. Springer, 1989.

[344] Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*, 2020.

[345] Douglas B Lenat. Eurisko: a program that learns new heuristics and domain concepts: the nature of heuristics iii: program design and results. *Artificial intelligence*, 21(1-2):61–98, 1983.

[346] John C Lere and Bruce R Gaumnitz. The impact of codes of ethics on decision making: Some insights from information economics. *Journal of Business Ethics*, 48(4):365–379, 2003.

[347] John-Clark Levin and Matthijs M Maas. Roadmap to a roadmap: How could we tell when agi is a'manhattan project'away? *arXiv preprint arXiv:2008.04701*, 2020.

[348] Jonathan Levin and Barry Nalebuff. An introduction to vote-counting schemes. *Journal of Economic Perspectives*, 9(1):3–26, 1995.

[349] David Lewis. *Counterfactuals*. John Wiley & Sons, 2013.

[350] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim

Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[351] Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation, 2023.

[352] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.

[353] Haoran Li, Yiran Liu, Xingxing Zhang, Wei Lu, and Furu Wei. Tuna: Instruction tuning using feedback from large language models. *arXiv preprint arXiv:2310.13385*, 2023.

[354] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[355] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *arXiv preprint arXiv:2203.07814*, 2022.

[356] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[357] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interven-

tions: explanation and elaboration. *Annals of internal medicine*, 151(4):W–65, 2009.

[358] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[359] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[360] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*, 2023.

[361] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019.

[362] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

[363] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.

[364] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[365] Daniel J Lizotte, Michael Bowling, and Susan A Murphy. Linear fitted-q

iteration with multiple reward functions. *The Journal of Machine Learning Research*, 13(1):3253–3295, 2012.

[366] Daniel J Lizotte, Michael H Bowling, and Susan A Murphy. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *ICML*, volume 10, pages 695–702, 2010.

[367] Seth Lloyd. Computational capacity of the universe. *Physical Review Letters*, 88(23):237901, 2002.

[368] John Locke. *The second treatise of civil government*. Broadview Press, 2015.

[369] Ted Lockhart. *Moral uncertainty and its consequences*. Oxford University Press, 2000.

[370] Thomas Lombardo. Consciousness, cosmic evolution, and the technological singularity. *Journal of Futures Studies*, 17(2):93–100, 2012.

[371] John R Lucas. Minds, machines, and gödel: A retrospect. *Machines and thought: The legacy of Alan Turing*, 1:103–124, 1996.

[372] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*, 2022.

[373] Yicheng Luo, Jackie Kay, Edward Grefenstette, and Marc Peter Deisenroth. Finetuning from offline reinforcement learning: Challenges, trade-offs and practical solutions. *arXiv preprint arXiv:2303.17396*, 2023.

[374] Richard Luscombe. Google engineer put on leave after saying ai chatbot has become sentient. *The Guard*, 2022.

[375] Matthijs M Maas. How viable is international arms control for military artificial intelligence? three lessons from nuclear weapons. *Contemporary Security Policy*, 40(3):285–311, 2019.

[376] Michael MacAskill, Krister Bykvist, and Toby Ord. *Moral uncertainty*. Oxford University Press, 2020.

[377] William MacAskill. *Normative uncertainty*. PhD thesis, University of Oxford, 2014.

[378] William MacAskill. Normative uncertainty as a voting problem. *Mind*, 125(500):967–1004, 2016.

[379] William MacAskill and Toby Ord. Why maximize expected choice-worthiness? *Noûs*, 54(2):327–353, 2020.

[380] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[381] Jocelyn Maclure. Context, intersubjectivism, and value: Humean constructivism revisited. *Dialogue: Canadian Philosophical Review/revue Canadienne De Philosophie*, 59(3):377–401, 2020.

[382] Victor Mardellat. Contractualism and the paradox of deontology. *Philosophical Studies*, 177(12):3749–3774, 2020.

[383] Eric Maskin. A modified version of arrow's iia condition. *Social Choice and Welfare*, 54(2-3):203–209, 2020.

[384] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. The ai index 2023 annual report. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University*, 2023.

[385] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on arti-*

ficial intelligence, volume 35, pages 14867–14875, 2021.

[386] Kenneth O May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica: Journal of the Econometric Society*, pages 680–684, 1952.

[387] Tim May. Cyphernomicon, 1994.

[388] Timothy May. The crypto anarchist manifesto. *High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace*, 1992.

[389] Ernst Mayr. Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506, 1961.

[390] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12, 2006.

[391] Pamela McCorduck and Cli Cfe. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press, 2004.

[392] Daniel McIntosh. We need to talk about data: how digital monopolies arise and why they have power and influence. *J. Tech. L. & Pol'y*, 23:185, 2018.

[393] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

[394] Christopher McMahon. The paradox of deontology. *Philosophy & Public Affairs*, pages 350–377, 1991.

[395] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. Does acm's code of ethics change ethical decision making in software development? In

Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, pages 729–733, 2018.

[396] William H Melody. Telecom reform: progress and prospects. *Telecommunications Policy*, 23(1):7–34, 1999.

[397] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*, 2022.

[398] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

[399] Maurice Merleau-Ponty. *Phenomenology of perception*. Routledge, 2013.

[400] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

[401] Steven H Miles. *The Hippocratic Oath and the ethics of medicine*. Oxford University Press, 2005.

[402] John Stuart Mill. Utilitarianism (1863). *Utilitarianism, Liberty, Representative Government*, pages 7–9, 1859.

[403] Ruth Garrett Millikan. *Language, thought, and other biological categories: New foundations for realism*. MIT press, 1987.

[404] The Villani Mission. National strategy for ai. Technical report, AI for humanity, 2018.

[405] Melanie Mitchell. Debates on the nature of artificial general intelligence, 2024.

[406] Brent Mittelstadt. Principles alone cannot guarantee ethical AI. 1(11):501–507, 11 2019.

[407] Jonas Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4:347–365, 1994.

[408] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[409] Dagmar Monett, Colin W. P. Lewis, Kristinn R. Thórisson, Joscha Bach, Gianluca Baldassarre, Giovanni Granato, Istvan S. N. Berkeley, François Chollet, Matthew Crosby, Henry Shevlin, John Fox, John E. Laird, Shane Legg, Peter Lindes, Tomáš Mikolov, William J. Rapaport, Raúl Rojas, Marek Rosa, Peter Stone, Richard S. Sutton, Roman V. Yampolskiy, Pei Wang, Roger Schank, Aaron Sloman, and Alan Winfield. Special issue "on defining artificial intelligence"—commentaries and author's response. 11(2):1–100, 2 2020.

[410] Dagmar Monett and Colin WP Lewis. Getting clarity by defining artificial intelligence—a survey. In *3rd conference on" philosophy and theory of artificial intelligence*, pages 212–214. Springer, 2017.

[411] Dena Montague. Stolen goods. *SAIS Review (1989-2003)*, 22(1):103–118, 2002.

[412] Beronda L Montgomery. Academic leadership: gatekeeping or groundskeeping? *The Journal of Values-Based Leadership*, 13(2):16, 2020.

[413] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022.

[414] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal,

Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

[415] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 462–478. Springer, 2023.

[416] Vincent C Müller and Nick Bostrom. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence*, pages 555–572. Springer, 2016.

[417] Luke Munn. The uselessness of ai ethics. *AI and Ethics*, pages 1–9, 2022.

[418] AO Mustapha, P Mbuzukongira, and MJ Mangala. Occupational radiation exposures of artisans mining columbite–tantalite in the eastern democratic republic of congo. *Journal of Radiological Protection*, 27(2):187, 2007.

[419] Kenjiro Nakamura. The vetoers in a simple game with ordinal preferences. *International Journal of Game Theory*, 8:55–61, 1979.

[420] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[421] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt (2019). *arXiv preprint arXiv:1912.02292*, 6, 2019.

[422] Jan H Naude. Technological singularity and transcendental monism: co-producers of sustainable alternative futures. *Journal of Futures Studies*, 13(3):49–58, 2009.

[423] John J Nay. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *arXiv preprint arXiv:2209.13020*, 2022.

[424] Michael Nest. *Coltan*, volume 3. Polity, 2011.

[425] Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994.

[426] Allen Newell and Herbert Alexander Simon. Gps, a program that simulates human thought. 1961.

[427] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

[428] Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

[429] Moti Nissani. Fruits, salads, and smoothies: A working definition of interdisciplinarity. *The Journal of Educational Thought (JET)/Revue de la Pensée Éducative*, pages 121–128, 1995.

[430] Robert Nozick. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

[431] Robert Nozick. *Anarchy, state, and utopia*, volume 5038. new york: Basic Books, 1974.

[432] University of Montréal. Montréal declaration for a responsible development of artificial intelligence. Technical report, University of Montréal, 2018.

[433] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread(June 27). http://www. transformer-circuits. pub/2022/mech-interp-essay/index. html*, 2022.

[434] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[435] IFD Oliveira, S Zehavi, and O Davidov. Stochastic transitivity: Axioms and models. *Journal of Mathematical Psychology*, 85:25–35, 2018.

[436] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[437] SM Omohundro. The basic ai drives-proceedings of the first agi conference. In *Frontiers in Artificial Intelligence and Applications*, 2008.

[438] AI HLEG High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. 2019.

[439] OpenAI. Chatgpt. `https://chat.openai.com/`, 2024.

[440] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[441] Laurent Orseau, Simon McGregor McGill, and Shane Legg. Agents and devices: A relative definition of agency. *arXiv preprint arXiv:1805.12387*, 2018.

[442] Mary Osborn, Richard Day, Paul Komesaroff, and Andrea Mant. Do ethical guidelines make a difference to decision-making? *Internal medicine journal*, 39(12):800–805, 2009.

[443] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, 2021.

[444] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[445] C O'Keefe, P Cihon, C Flynn, B Garfinkel, J Leung, and A Dafoe. The windfall clause: Distributing the benefits of ai. centre for the governance of ai research report. future of humanity institute, university of oxford, 2020.

[446] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

[447] Nicolas Papernot. A marauder's map of security and privacy in machine learning: An overview of current and future research directions for making machine learning secure and private. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 1–1, 2018.

[448] Vivak Patel and Shushu Zhang. Stochastic gradient descent on nonconvex functions with general noise models. *arXiv preprint arXiv:2104.00423*, 2021.

[449] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

[450] Roger Penrose. *Shadows of the Mind*, volume 4. Oxford University Press Oxford, 1994.

[451] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[452] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

[453] Billy Perrigo. Exclusive: Openai used kenyan workers on less than $2 per hour to make chatgpt less toxic. *Time*, 2023.

[454] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[455] Michael A Peters. Technological unemployment: Educating for the fourth industrial revolution. In *The Chinese Dream: Educating the Future*, pages 99–107. Routledge, 2019.

[456] Nicolas Petit. Big tech and the digital economy: The moligopoly scenario. 2020.

[457] Jean Piaget. O nascimento da inteligência na criança. *mental*, 258:259, 1986.

[458] Jean Piaget. *O juízo moral na criança*. Grupo Editorial Summus, 1994.

[459] Gualtiero Piccinini. The mind as neural software? understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, 81(2):269–311, 2010.

[460] Jeffrey Stephen Poland. Physicalism, the philosophical foundations. 1994.

[461] Martha E Pollack. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI magazine*, 26(2):9–9, 2005.

[462] Boris Teodorovich Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[463] Simon JD Prince. *Understanding Deep Learning*. MIT press, 2023.

[464] Hilary Putnam. Review of shadows of the mind. *AMS Bulletin*, 32(3), 1995.

[465] Hilary Putnam et al. *Reason, truth and history*, volume 3. Cambridge University Press, 1981.

[466] PricewaterhouseCoopers (PwC). A practical guide to responsible artificial intelligence, 2019.

[467] Zhuwei Qin, Fuxun Yu, Chenchen Liu, and Xiang Chen. How convolutional neural network see the world-a survey of convolutional neural network visualization methods. *arXiv preprint arXiv:1804.11191*, 2018.

[468] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[469] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[470] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[471] Peter Railton. Facts and values. *Philosophical Topics*, 14(2):5–31, 1986.

[472] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[473] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

[474] Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 6383–6402, 2023.

[475] John Rawls. A theory of justice. In *Ethics*, pages 229–234. Routledge, 2004.

[476] John Rawls et al. *The law of peoples*. na, 1996.

[477] Fred Reed. Promise of ai not so bright. *The Washington Times*, 2006.

[478] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[479] Anaïs Rességuier and Rowena Rodrigues. <i>ai ethics should not remain toothless!</i> A call to bring back the teeth of ethics. 7(2):205395172094254, 7 2020.

[480] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[481] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[482] Arthur Ripstein. Beyond the harm principle. *Philosophy & Public Affairs*, 34(3):215–245, 2006.

[483] Howard Robinson. Dualism. 2003.

[484] W Teed Rockwell. *Neither brain nor ghost: A nondualist alternative to the mind-brain identity theory*. Mit Press Cambridge, MA, 2005.

[485] Heather M Roff. The frame problem: The ai "arms race" isn't one. *Bulletin of the Atomic Scientists*, 75(3):95–98, 2019.

[486] Hartmut Rosa and William E. Scheuerman. *High-Speed Society: Social Acceleration, Power, and Modernity*. 2009.

[487] Connie S Rosati. Persons, perspectives, and full information accounts of the good. *Ethics*, 105(2):296–325, 1995.

[488] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[489] Stanley J Rosenschein and Leslie Pack Kaelbling. The synthesis of digital machines with provable epistemic properties. In *Theoretical aspects of reasoning about knowledge*, pages 83–98. Elsevier, 1986.

[490] Jacob Ross. Rejecting ethical deflationism. *Ethics*, 116(4):742–768, 2006.

[491] Corby Rosset. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog*, 1(2), 2020.

[492] Jörg Rothe et al. *Economics and computation*, volume 4. Springer, 2015.

[493] David Rotman. How technology is destroying jobs. *Technology Review*, 16(4):28–35, 2013.

[494] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*, 2022.

[495] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*, 2020.

[496] Jean-Jacques Rousseau. *The social contract: or, principles of political right*. Number 83. G. Allen & Unwin, Limited, 1916.

[497] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code

llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[498] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[499] Stuart Russell. Of myths and moonshine. *Retrieved from edge. org/conversation/jaron_lanier-the-myth-of-ai*, 2014.

[500] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[501] Stuart Russell. Ai weapons: Russia's war in ukraine shows why the world must enact a ban. *Nature*, 614(7949):620–623, 2023.

[502] Stuart Russell, Daniel Dewey, and Max Tegmark. An open letter: Research priorities for robust and beneficial artificial intelligence. open letter. signed by 8,600 people, 2015.

[503] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[504] Gilbert Ryle. *The concept of mind*. Routledge, 2009.

[505] Donald Saari. *Decisions and elections: explaining the unexpected*. Cambridge University Press, 2001.

[506] Donald G Saari and Vincent R Merlin. The copeland method: I.: relationships and the dictionary. *Economic theory*, 8:51–76, 1996.

[507] Laura Sallstrom, Olive Morris, and Halak Mehta. Artificial intelligence in africa's healthcare: Ethical considerations, 2019.

[508] I Sample. Science journals ban listing of chatgpt as co-author on papers. *The Guardian*, 26, 2023.

[509] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and

Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *ACL*, 2020.

[510] Faridun Sattarov. *Power and technology: A philosophical and ethical analysis*. Rowman & Littlefield, 2019.

[511] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

[512] Neil Savage. The race to the top among the world's leaders in artificial intelligence. 588(7837):S102–S104, 12 2020.

[513] Geoffrey Sayre-McCord. Coherentist epistemology and moral theory. 1996.

[514] Thomas M Scanlon. *What we owe to each other*. Harvard University Press, 2000.

[515] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[516] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. What's next for ai ethics, policy, and governance? a global overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 153–158, 2020.

[517] Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas. Ai ethics in the public, private, and ngo sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1):31–42, 2021.

[518] Jürgen Schmidhuber. Algorithmic theories of everything. *arXiv preprint quant-ph/0011122*, 2000.

[519] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neu-*

ral networks, 61:85–117, 2015.

[520] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[521] Susan Schneider. How to catch an ai zombie. *Ethics of Artificial Intelligence*, page 439, 2020.

[522] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[523] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[524] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social choice and Welfare*, 36(2):267–303, 2011.

[525] Markus Schulze. The schulze method of voting. *arXiv preprint arXiv:1804.02973*, 2018.

[526] Robert E Scott and George G Triantis. Incomplete contracts and the theory of contract design. *Case W. Res. L. Rev.*, 56:187, 2005.

[527] John Searle. Can computers think. *Minds, Brains, and Science*, pages 28–41, 1984.

[528] JR Searle. Minds, brains, and programs. behavioral and brain sciences3: 41724.[asl](1992) the rediscovery of mind, 1980.

[529] Nigel Seel. *Agent theories and architectures*. University of Surrey (United

Kingdom), 1989.

[530] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[531] Amartya Sen. Behaviour and the concept of preference. *Economica*, 40(159):241–259, 1973.

[532] Amartya Sen. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181, 1986.

[533] Amartya Sen. *Rationality and freedom*. Harvard University Press, 2004.

[534] Amartya Sen. *Collective choice and social welfare: An expanded edition*. Harvard University Press, 2017.

[535] Amartya K Sen. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, pages 317–344, 1977.

[536] Russ Shafer-Landau. *Moral realism: A defence*. Clarendon Press, 2003.

[537] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. Preferences implicit in the state of the world. *arXiv preprint arXiv:1902.04198*, 2019.

[538] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[539] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[540] Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring se-

curity vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*, 2023.

[541] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

[542] Yoav Shoham. Agent-oriented programming. *Artificial intelligence*, 60(1):51–92, 1993.

[543] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2659–2673, 2022.

[544] Henry Sidgwick. *The methods of ethics*. Good Press, 2019.

[545] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[546] Walter Sinnott-Armstrong. *Moral skepticisms*. Oxford University Press, Oxford; New York, 2006.

[547] Linda B Smith and Esther Thelen. Development as a dynamic system. *Trends in cognitive sciences*, 7(8):343–348, 2003.

[548] Nate Soares. The value learning problem. In *Artificial Intelligence Safety and Security*, pages 89–97. Chapman and Hall/CRC, 7 2018.

[549] Nate Soares and Benja Fallenstein. Toward idealized decision theory. *arXiv preprint arXiv:1507.01986*, 2015.

[550] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[551] Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

[552] Ernest Sosa et al. *Knowledge in perspective: Selected essays in epistemology*. Cambridge university press, 1991.

[553] Kaj Sotala. *Artificial Intelligence Safety and Security*. 2018.

[554] Kaj Sotala. Disjunctive scenarios of catastrophic ai risk. In *Artificial Intelligence Safety and Security*, pages 315–337. Chapman and Hall/CRC, 2018.

[555] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[556] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[557] Robert C Stalnaker. A theory of conditionals. In *Ifs*, pages 41–55. Springer, 1968.

[558] Eric Steinhart. Digital metaphysics. 1998.

[559] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Process-*

ing Systems, 33:3008–3021, 2020.

[560] Deborah D Stine. The manhattan project, the apollo program, and federal energy technology r & d programs: A comparative analysis. Congressional Research Service, the Library of Congress, 2008.

[561] Daniel Stoljar. Emotivism and truth conditions. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 70(1):81–101, 1993.

[562] Daniel Stoljar. *Physicalism*. Routledge, 2010.

[563] Sharon Street. Coming to terms with contingency: Humean constructivism about practical reason. *Constructivism in practical philosophy*, pages 40–59, 2012.

[564] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[565] Cass R Sunstein. Beyond the precautionary principle. *U. Pa. L. Rev.*, 151:1003, 2002.

[566] Ewan Sutherland. Coltan, the congo and your cell phone. *Available at SSRN 1752822*, 2011.

[567] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019.

[568] Robert Tally. *New Essays on the Frankfurt School of Critical Theory*. 2009.

[569] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[570] Max Tegmark. Benefits and risks of artificial intelligence. *Future of life*, pages 29–31, 2016.

[571] Max Tegmark. *Life 3.0: Being human in the age of artificial intelligence*. Vintage, 2017.

[572] Max Tegmark. Max tegmark: The case for halting ai development | lex fridman podcast #371, Aug. 2019.

[573] Tencent Research Institute (Tencent). ""arcc"": An ethical framework for artificial intelligence, 2018.

[574] Esther Thelen and Linda B Smith. *A dynamic systems approach to the development of cognition and action*. MIT press, 1994.

[575] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[576] Nicolaus Tideman. *Collective decisions and voting: the potential for public choice*. Ashgate Publishing, Ltd., 2006.

[577] T Nicolaus Tideman. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206, 1987.

[578] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

[579] Phil Torres. Superintelligence and the future of governance: On prioritizing the control problem at the end of history. In *Artificial intelligence safety and security*, pages 357–374. Chapman and Hall/CRC, 2018.

[580] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[581] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[582] BJ Trawinski and HA David. Selection of the best treatment in a paired-comparison experiment. *The Annals of Mathematical Statistics*, 34(1):75–91, 1963.

[583] Joseph Trevithick. Turkey now has swarming suicide drones it could export. *The Drive*, 2020.

[584] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[585] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524, 2018.

[586] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[587] Alan Mathison Turing et al. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.

[588] Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. Avoiding side effects in complex environments. *Advances in Neural Information Processing Systems*, 33:21406–21415, 2020.

[589] Alexander Matt Turner. On avoiding power-seeking by artificial intelligence. *arXiv preprint arXiv:2206.11831*, 2022.

[590] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Con-

servative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

[591] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. *arXiv preprint arXiv:1912.01683*, 2019.

[592] Piers Norris Turner. "harm" and mill's harm principle. *Ethics*, 124(2):299–326, 2014.

[593] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. In *Behavioral decision making*, pages 25–41. Springer, 1985.

[594] Bibek Upadhayay and Vahid Behzadan. Sandwich attack: Multi-language mixture adaptive attack on llms. *arXiv preprint arXiv:2404.07242*, 2024.

[595] Artur Usanov, Marjolein de Ridder, Willem Auping, Stephanie Lingemann, Luis Tercero Espinoza, Magnus Ericsson, and M Liedtke. *Coltan, Congo & conflict: Polinares case study*. JSTOR, 2013.

[596] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. "this is just a prototype": How ethics are ignored in software startup-like environments. In *International Conference on Agile Software Development*, pages 195–210. Springer, Cham, 2020.

[597] Merijn Van Erp, Louis Vuurpijl, and Lambert Schomaker. An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 195–200. IEEE, 2002.

[598] Tim Van Gelder. The dynamical hypothesis in cognitive science. *Behavioral and brain sciences*, 21(5):615–628, 1998.

[599] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[600] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[601] David Vernon. Robotics and artificial intelligence in africa [regional]. 26(4):131–135, 12 2019.

[602] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*, 2020.

[603] James Vincent. Microsoft's bing is an emotionally manipulative liar, and people love it. *The Verge*, 15, 2023.

[604] Vernor Vinge. The coming technological singularity: How to survive in the post-human era. *Science fiction criticism: An anthology of essential writings*, pages 352–363, 1993.

[605] Paul Virilio. *Open Sky*. 1997.

[606] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

[607] Rosalie Waelen. Why ai ethics is a critical theory. *Philosophy & Technology*, 35(1):1–16, 2022.

[608] Wendell Wallach and Shannon Vallor. Moral machines. *Ethics of Artificial Intelligence. Oxford University Press*, pages 383–412, 2020.

[609] Denis M Walsh. Teleology. 2008.

[610] Pei Wang. *Non-axiomatic reasoning system: Exploring the essence of intelligence*. Indiana University, 1995.

[611] Pei Wang. What do you mean by" ai"? In *AGI*, volume 171, pages 362–373,

2008.

[612] Pei Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019.

[613] Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. *arXiv preprint arXiv:2310.13671*, 2023.

[614] Yanchen Wang and Lisa Singh. Adding guardrails to advanced chatbots. *arXiv preprint arXiv:2306.07500*, 2023.

[615] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[616] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[617] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[618] John Welsh. Tolling academics: Rent-seeking and gatekeeping in the university space. *Capital & Class*, 45(1):93–121, 2021.

[619] Hermann Weyl. The open world: Three lectures on the metaphysical implications of science. 1932.

[620] John Archibald Wheeler. *Information, physics, quantum: The search for links*. CRC Press, 2018.

[621] Edwin B Wilson. The abilities of man, their nature and measurement. by

c. spearman. new york, the macmillan co., 1927. vi+ 415+ xxxii pp. *Science*, 67(1731):244–248, 1928.

[622] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12, 2017.

[623] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[624] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

[625] WolframAlpha. Wolframalpha.

[626] David B Wong. *Natural moralities: A defense of pluralistic relativism*. Oxford University Press, 2006.

[627] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[628] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the*

2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2950–2968, 2021.

[629] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[630] Hiroshi Yamakawa. Peacekeeping conditions for an artificial intelligence society. *Big Data and Cognitive Computing*, 3(2):34, 2019.

[631] Roman V Yampolskiy. Leakproofing singularity-artificial intelligence confinement problem. *Journal of Consciousness Studies JCS*, 2012.

[632] Roman V Yampolskiy. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and theory of artificial intelligence*, pages 389–396. Springer, 2013.

[633] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[634] Karen Yeung. Recommendation of the council on artificial intelligence (OECD). 59(1):27–34, 2 2020.

[635] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[636] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[637] Eliezer Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.

[638] Eliezer Yudkowsky. Timeless decision theory. *The Singularity Institute, San Francisco*, 2010.

[639] Eliezer Yudkowsky. Complex value systems in friendly ai. In *International Conference on Artificial General Intelligence*, pages 388–393. Springer, 2011.

[640] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016.

[641] Eliezer Yudkowsky et al. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303):184, 2008.

[642] Eliezer Yudkowsky and Nate Soares. Functional decision theory: a new theory of instrumental rationality. *arXiv preprint arXiv:1710.05060*, 2017.

[643] Lotfi A Zadeh. A computational theory of dispositions. In *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*, pages 713–737. World Scientific, 1996.

[644] Ernst Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.

[645] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. arxiv e-prints, page. *arXiv preprint arXiv:1611.03530*, 2016.

[646] Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. The AI index 2022 annual report. 05 2022.

[647] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. The ai index 2021 annual report. *arXiv preprint*

arXiv:2103.06312, 2021.

[648] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

[649] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[650] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[651] Mandi Zhao, Pieter Abbeel, and Stephen James. On the effectiveness of fine-tuning versus meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 35:26519–26531, 2022.

[652] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena, arxiv abs/2306.05685 (2023). *URL: https://api. semanticscholar. org/CorpusID*, 259129398, 2023.

[653] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

[654] Jichen Zhu and D Fox Harrell. System intentionality and the artificial intelligence hermeneutic network: the role of intentional vocabulary. In *Proceedings of the 2009 Digital Art and Culture Conference*, 2009.

[655] Daniel M Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter

Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, et al. Adversarial training for high-stakes reliability. *arXiv preprint arXiv:2205.01663*, 2022.

[656] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[657] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[658] Konrad Zuse. Rechnender raum (calculating space). *Schriften Zur Dataverarbeitung*, 1, 1969.

MARISTA PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br