Benevolent Guardians or Strategic Actors? A Critical Analysis of Corporate Governance Legitimation in Leading Artificial Intelligence Labs

By Samuel Archer

Word count: 14,992

Abstract

This paper critically analyses the discursive legitimation strategies used by leading AI companies, in their legitimation of their novel and untested corporate governance structures. Using a critical discourse analysis (CDA), that combines the methodologies of van Leeuwen (2007) and Fairclough (2010), this study analysed official texts published on AI corporations' websites and transcripts of podcast interviews with their respective CEOs. Results established persistent and strategic attempts by each of the AI companies to legitimate their corporate governance structures, using a variety of strategies to do so.

In doing so it yields insights into the complex links between AI safety, legitimation and corporate governance, through facilitating a better understanding of strategies used and their wider implications. The dissertation contributes to the respective bodies of literature on AI safety, corporate governance and legitimation, representing the first study to bring together these three areas and providing novel empirical insights into the legitimation processes of a rapidly evolving industry. It also offers practical recommendations, both for industries seeking to legitimate their actions, and policymakers seeking to guarantee the development of safe and responsible AI.

Table of Contents

Chapter 1: Introduction 7

1.1 Aims of the research and central questions 8

1.2 Research Approach and Key Findings & Contributions 9

1.3 Structure of the Dissertation 10

Chapter 2: Literature Review 12

2.1 Artificial Intelligence Safety 12

2.1.1 Introducing Artificial Intelligence Safety 12

2.1.2 Policy and Regulation AI Safety Research 13

2.1.3 Corporate AI Safety Approaches 14

2.2 Corporate Governance 15

2.2.1 Dominant Theories of Corporate Governance 15

2.2.2 The Fundamentals of Corporate Governance 16

2.2.3 Corporate Social Responsibility (CSR) 17

2.3 Legitimacy and Legitimation 18

2.3.1 Introducing Legitimacy 18

2.3.2 Legitimation Differentiated from Legitimacy 19

2.3.3 Legitimation Theory 20

2.4 Concluding the Literature Review 21

Chapter 3: Methodology 22

3.1 Research Philosophy 22

3.2 Design of the Study 23

3.3 Data Collection 24

3.4 Data Analysis 25

3.4.1 CDA Design 25

3.4.2 CDA Implementation 27

3.5 Approach to Theorising 30

3.6 Ethical Considerations and Limitations 30
3.7 Summary 31
Chapter 4: Findings, Analysis, and Discussion 32
4.1 Dialectical Relationship Analysis 32
4.2 Contextual (Socio-Corporate) Analysis 33
4.3 Intertextual Analysis 37
4.3.1 Militaristic Order of Discourse 37
4.3.2 Governmental Order of Discourse 39
4.3.3 Scientific and Academic Orders of Discourse 40
4.4 Textual Analysis 43
4.4.1 Ubiquitous Authorisation 43
4.4.2 Company-Line Moralisation 48
4.4.3 CEOs' Hero-Driven Narratives and Cautionary Tales 51
4.4.4 Rationalization Abound 53
Chapter 5: Conclusion 59
5.1 Answering the Research Questions 59
5.2 Contributions, Recommendations, and Future Research Directions 60
5.3 Final Remarks 61
Bibliography 62

List of Figures

Figure 1: Ouroborous, or The Endless Endeavor of Legitimation 19
Figure 2: Visual representation of the phases of Analysis, based on Fairclough's (2010) CDA methodology 27
Figure 3: Visual representation of OpenAI's corporate governance structure (Figure taken from (OpenAI, 2024)) 34
Figure 4: Anthropic's corporate governance structure (Figure created by author, based upon Text 2) 35

List of Tables

Table 1: Selection of Texts for Analysis 24
Table 2: Summary of van Leeuwen's Categories of Legitimation 29
Table 3: Militaristic (and Informal) Orders of Discourse 37
Table 4: Governmental Order of Discourse 39
Table 5: Scientific and Academic Orders of Discourse 41
Table 6: Varying Forms of Authorisation 45
Table 7: 'Temporal' Authorisation 46
Table 8: Company-Line Moralisation 49
Table 9: CEO Mythopoeia 52
Table 10: Theoretical Rationalisations 54
Table 11: Instrumental Rationalisation 56

Table of Abbreviations

CG	Corporate Governance
AI	Artificial Intelligence
AGI	Aritificial General Intelligence
CDA	Critical Discourse Analysis
CSR	Corporate Social Responsibility
LTBT	Long-Term Benefit Trust
PBC	Public Benefit Corporation

Acknowledgements

Without the enduring and invaluable guidance, wisdom and support provided by my dissertation supervisor, Professor Stefanie Reissner, this dissertation would look very different from how it does today. As such, I would like to extend to her my deepest and most sincere gratitude.

I would like also to thank my family, for their boundless support and love, throughout not only the writing of this dissertation but in all that I do.

Chapter 1: Introduction

"The road to AGI should be a giant power struggle."

Sam Altman (2024), CEO and Board Member of OpenAI

Talk of artificial general intelligence (AGI) has percolated in the computer science literature since the advent of computers in the 1940s. However, the past ten years have seen developments that suggest this technology will not long be restricted to the realm of science fiction (Juric et al., 2020). Radical leaps in machine learning and deep learning, coupled with forecasted increases in compute power (Theis & Wong, 2017) and compute efficiency (Kaplan, et al., 2020), have led to the majority of academic AI researchers believing there to be a 50% chance of AI bettering humans in all tasks by 2063 (Grace et al., 2018). Such an impact would mark a technological leap akin to the industrial and agricultural revolutions, profoundly affecting our societal and business landscapes (Makridakis, 2017).

Whilst such developments may bring untold benefits, AI's proliferation also brings with it a multitude of potential dangers, ranging from discrimination to humanity-wide existential threats (Mouton et al., 2024; Sotala, 2018). Whoever develops these technologies thus inherits significant responsibility. At present, it seems probable that the world's frontier AI models will continue to be developed by corporations, with the industry's progress far outpacing both academia and governments (HAI, 2024). Current incarnations of this technology are already impacting the corporate landscape, as indicated by ChatGPT's 2022 release marking the fastest-ever growth of a consumer application, with multiple reports anticipating near-future AI-caused job displacement (McKinsey, 2023) and economy-wide industry disruption (HAI, 2024). The world's leading AI labs may therefore soon maintain a level of power previously seen only by governments, a possibility made greater by the billions of dollars of funding received from the world's largest corporations (Amazon, Google, Microsoft, and Meta),

Current AI safety measures from the policy and regulatory side are fundamentally limited [See: Section 2.1.2], meaning internal corporate governance structures in these leading labs play a

vital role in tempering the near-monopolistic power and control that AI corporations maintain. Whilst "[c]orporate governance power can be used to achieve outstanding results", however, it can also be abused, meaning that the "politics and ethics of power are intrinsic to corporate governance" (Tricker, 2020, p. 58). As several leading AI labs have developed and implemented novel and untested corporate governance structures, these concerns are exacerbated.

Language represents a core manifestation and realisation of such power, with it doing extensive "social and ideological 'work'" in "producing, reproducing, or transforming social structures, relations and identities" (Fairclough, 1992, p. 211)**.** Thereby, equal to the need for strong corporate governance structures within AI labs is a need for an effective, well-grounded and thorough criticism of the corporate governance structures these companies maintain. A fundamental tool of discursive power manipulation is that of legitimation, meaning how a social actor justifies their existence and activities through language (van Leeuwen, 2007). The acquisition and maintenance of legitimacy is essential for all corporations (Berger & Luckmann, 1967, p. 82) but industries early in their development, such as AI, encounter legitimacy issues due to a lack of historical precedence (Zimmerman & Zeitz, 2002). As such, not only do AI companies have to "devote a substantial amount of energy to sector building" (Suchman, 1995, p. 586) to establish legitimacy for themselves and their industry, but they are also forced to legitimate their corporate governance structures through discourse and varying legitimation techniques. Such strong incentives to legitimate, coupled with these labs' significant power and recent corporate governance crises [See: Section 4.2], necessitate rigorous critical analysis.

1.1 Aims of the research and central questions

On its broadest level, this study seeks to contribute to the central question of AI safety, which asks 'How can we build AI in a way that is both safe and beneficial for all of humanity?'. As is argued in the Literature Review [Section 2], the management literature's direct contribution to this topic is so far lacking. Focusing its attention on the three central themes of corporate governance, AI safety, and legitimacy, this study will endeavour to analyse the patterns of how leading AI labs are discursively legitimating their novel and untested corporate governance structures. This research's three research questions can be summarised as follows:

RQ1: What are the textual and intertextual legitimation techniques that leading AI labs use to legitimate their novel corporate governance structures?

RQ2: How do leading AI labs conceptualise their responsibilities in relation to themselves and corporate governance?

RQ3: What wider implications do these findings have for AI safety?

The aims of this study are not to evaluate the corporate governance structures themselves, nor the constituencies of their boards. Indeed, it will also not seek to address their respective capacities to deal with threats to safety. Instead, it is focused on understanding how power is being used to shape perceptions and realities using legitimative discourse. In doing so, it is hoped that this will improve understanding of the intentions, practices and ambitions of AI labs, which will in turn assist in procuring greater standards and norms around AI safety, thus ultimately benefiting both industry and humanity.

1.2 Research Approach and Key Findings & Contributions

In answering the three research questions, this research utilised a documentary study of naturally occurring data, in the form of texts published on the websites of two leading AI labs [Appendix 1 & 2] and transcripts of podcast interviews with their respective CEOs [Appendix 3 & 4]. Throughout the subsequent analysis, a critical perspective was utilised, to enable a deeper interpretation of the discursive moves and their implicit intentions, as well as a comparison between the labs, their CEOs, and their internal contradictions. For analysing the data, Fairclough's three-dimensional critical discourse analysis (CDA) methodology was utilised, to facilitate comprehensive discursive, contextual and intertextual analyses, whilst van Leeuwen's legitimation CDA approach was employed for the close-point textual analysis. In developing a coding schema for the textual analysis [4.4], van Leeuwen's categories of legitimation were used to retain greater theoretical consistency.

This dissertation yields several significant findings and makes notable contributions to the field. Foremost among these is this research's contribution in representing the first-ever study

linking the disciplines of AI safety and legitimation, as well as contributing to the presently underserved yet deeply important literature on AI safety and corporate governance. Significant findings centre around the persistent and systematic attempt by leading AI corporations to discursively legitimate their corporate governance structures. Almost all of van Leeuwen's (2007) legitimation techniques were evidenced, with authorisation and rationalisation used ubiquitously by all actors, including a strategy of legitimation ('temporal authorisation') not previously discussed in the legitimation literature.

This study also found that corporations' official publications favoured heavy moral evaluation [Section 4.4.2], using also militaristic, governmental and scientific orders of discourse to further legitimate themselves [Sections 4.3], whilst their CEOs made greater use of mythopoeia, especially in instances of previously diminished legitimacy. In doing so, this study has made contributions to both the corporate governance and legitimation literature, by finding evidence supporting previous studies' results and making minor novel contributions to legitimation theory [Section 4.4.2] as well as the methodological approach [Section 3.4.1].

1.3 Structure of the Dissertation

The structure of this dissertation is intended to reflect the best means by which to address the core aims.

Having outlined in the Introduction [Chapter 1] the necessity for this research, as well as its key themes, its intentions and its relevance to both the management literature and industry, attention is subsequently turned to the Literature Review [Chapter 2]. Within this chapter, the academic literature relevant to this research is considered and evaluated, with gaps and strengths accordingly highlighted. The Literature review is structured into three main sections that align directly with the study's central themes, of AI safety [Section 2.1], corporate governance [Section 2.2], and legitimacy and legitimation [Section 2.3], with further sub-sections providing greater depth and observation.

The subsequent chapter [Chapter 3: Methodology] concerns itself with outlining the philosophies and processes that underpin and facilitate the analysis of the data. This chapter moves along the figurative lines of a whirlpool, beginning with the more abstract discussions of Research Philosophy [3.2] before spiralling down into sections on Data Collection [3.3] and Data Analysis [3.4], where this study's central analytical method in the form of a Critical Discourse Analysis (CDA) is introduced and explained. Throughout the chapter, each section is justified by its

philosophical underpinnings, which continue through considerations of the Theoretical Approach [3.5] and the Ethical and Limitational considerations [3.6].

This study's Findings, Analysis and Discussion can be found in Chapter 4. The structure of this chapter follows Fairclough's approach to CDA, first considering Dialectical Relationships [4.1], followed by the Contextual (Socio-Corporate) [4.2], Intertextual [4.3] and Textual [4.4] analyses. Whilst the exactitudes of each section's structure differ, owing to methodological and analytical parameters, similar patterns can be found through each. Within Sections 4.1 and 4.2, the findings, analysis, and discussion are interwoven throughout, whilst Sections 4.3 and 4.4 (and all their respective sub-sections) use summary tables to present core examples of each legitimation technique, then combine the Analysis and Discussion throughout the accompanying text. Whilst the Discussion happens throughout this chapter, by situating the Findings in the wider research and expanding on the implications for AI safety, each sub-section within 4.3 and 4.4 contains a summarising discussion paragraph at its end.

Finally, in Chapter 5, this project is concluded. Coverage included is the summarising of answers to the core research aims, suggestions for further research as related to this study's findings, and an account of contributions that this study makes to both the wider literature and industry at large.

Chapter 2: Literature Review

As outlined in the Introduction [1.2], this chapter is structured to chart a path through the fruits of the academic literature that this study is focused upon: AI safety, corporate governance, and legitimacy & legitimation. AI safety will be first discussed, including an introduction of AI's core concepts of AI and an appraisal of work being done within AI safety, both corporate-oriented and otherwise. Having done so, attention will be turned to corporate governance [2.2], first looking at dominant theories [2.2.1], followed by an overview of the literature on relevant corporate governance structures [2.2.2], and the CG literature relating to CSR, with which AI safety shares many parallels. Indeed, the intention of this Section is not to provide a full account of the CG literature, but instead to introduce essential concepts to this study. Finally, in Section 2.3, the literature on legitimation and legitimacy will be evaluated, defining and differentiating these concepts [2.3.1 and 2.3.2], before zeroing in upon legitimation theory and its application in the management literature [2.3.3].

Whilst the literature centres on the management literature, my perspective is aligned with many key AI philosophers and scholars (Bostrom, 2014; Juric et al., 2020) in believing that 'solving' the issue of advanced AI safety will require a truly interdisciplinary approach. As such, some reference in the review is made to leading AI and legal journals, as well as some reports from leading corporate and journalistic sources, particularly in Section 2.1.2. Due to the necessarily limited scope of this project, however, as well as the strong disciplinary focus, the gaze of this review will rest predominantly upon the management literature.

2.1 Artificial Intelligence Safety

2.1.1 Introducing Artificial Intelligence Safety

Artificial intelligence is defined across varying literatures in differing ways, but Russell's (2022, p. 19) definition of non-biological actors that can "compute how to act effectively […] in a wide variety of novel situations" will be used, as will Tegmark's (2018, pp. 25-26) definition of AGI, as systems with the ability to "accomplish virtually any goal, including leaning". As argued in the opening paragraphs of the Introduction, the development of AI and AGIs could have potentially disastrous effects for humanity, as an increasingly impressive AI, computer science, and

philosophical literature has shown. These range from Bostrom's (2012) orthogonality thesis and Tegmark's (2018) "Life 3.0", to further work charting the dangerous (Sotala, 2018) and even potentially existential threats (Barrat, 2023; Bostrom, 2014) that AI may bring. As such, AI safety, defined as developing AI in ways that are "both safe and beneficial for humans" (Juric et al., 2020), is a field of great importance and contemporary relevance.

2.1.2 Policy and Regulation AI Safety Research

At present, technical approaches have accounted for the bulwark of the AI safety literature (Critch & Krueger, 2020), but the regulation and policy literature has also seen extensive advancements (Leslie, 2019). In addition to specific legal policy recommendations (Johnston, 2023), scholars have recommended the implementation of international AI standards (Cihon, 2019) and increased cooperation on AI to create strong economic, legal and domain-specific incentives for high safety standards (Askell et al., 2019, p. 16). Despite multiple scholars arguing that centralised regulation may be too "slow and brittle" to meaningfully make AI safe (Cihon et al., 2020, p. 228), regulation has begun to be implemented with the recently enacted regulatory-leading EU Artificial Intelligence Act (2024) marking a promising step towards AI regulation (Cihon et al., 2021, p. 2).

The current shortcomings of regulatory approaches have, however, provoked scholars to utilise research and progress made in the management literature to improve regulatory and governmental procedures. This has included calls to create a market for private, independent regulators to properly incentivize better regulation (Clark, 2019) using government-issued "bounties" (Love & Hubbard, 2009), and other calls to align AI regulation to socio-technological change, as was seen during the development of the internet and social media platforms (Maas, 2022). Further research has drawn inspiration from the management literature when proposing the mandating of process management certifications and AI certification procedures (Cihon et al., 2021, p. 200), as well as proposing a system focused on the implementing of independent assessment, audits, and adherence (Falco, et al., 2021).

Whilst the development of safe AI will no doubt include regulation and cooperation, there exists a convincing case made, in both the management and philosophy literatures, that such approaches alone will not suffice. Scholars have highlighted the impossibility of creating catch-all regulation (Yampolskiy, 2013), the ineffectualness of external risk assessment within corporations (Latin, 1988), and the extensive current lack of concrete policy recommendations in the field. The disastrous effects of one bad actor possessing one dangerous model, coupled with the novelty of

the subject area and the speed at which the technology is developing (Juric et al., 2020), result in such gaps in AI regulation being intolerable.

2.1.3 Corporate AI Safety Approaches

Coupling the inadequacy of the legal and regulatory safeguards around the development of advanced AI with corporations' uniquely powerful role in the development of these systems (HAI, 2024), leads to the conclusion that corporate governance has a significant role to play within AI safety. Developing AI safely is, as argued by Dubber (2020, p. 676), an "ongoing process that begins before product development and continues through product disposal", with a related concern here being the deeply advantageous nature of first-mover advantages within the technology sector (Lieberman & Montgomery, 1988). Recent reporting on AI has highlighted the emergence of an AI "arms race" (Roose, 2023), with Armstrong et al.'s (2016) Nash-equilibrium model of such a race anticipating dangerous consequences, including a depletion of quality control and lacking internal regulatory checks.

An attendant management literature here is the small, but growing, focus upon the corporate governance of emerging technologies. Marchant and Wallace (2013) explore frameworks for meta-level oversight of governance approaches and policies, highlighting the unique challenges that rapidly developing emerging technologies pose to traditional governance structures. However, their work lacks concrete proposals, instead advocating for the values of "adaptability" and "inclusivity", which are arguably insubstantial for addressing the central issues of AI safety. Within the AI safety literature there exists work highlighting the primacy of self-regulation by industry (LaGrandeur, 2021) as well as Cihon et al.'s (2021) paper, which isolates several actor-specific opportunities for AI governance, including for corporate partners, investors, workers and management. Whilst an excellent primer on the subject, the broadness of his brushstrokes and lack of concrete ideas do not represent the detailed approach necessary here.

In reviewing the management literature written directly about AI safety, its contribution to the subject is best described as disappointing, with it currently maintaining few meaningful contributions. A pre-print published by Tallarita (2023) represents an encouraging contribution to the management literature, in which he highlights the reasons and means by which AI is already testing the limits of corporate governance structures, including difficulties in taming profit motives and traditional corporate governance structures not being optimised to propagate wider public benefit. Given the perceived importance of addressing AI safety and corporations' power, this gap

in the management literature is perceived to be notable, with the need to fill it being both urgent and necessary.

2.2 Corporate Governance

Contrarily, the corporate governance literature is replete with major theoretical contributions. Since the conception of joint-stock limited liability companies in the 19th century, corporate governance has been a necessary force (Tricker, 2020) but it is only in the last three decades that it has "arrived" as a dominant paradigm in the management literature (Cheffins, 2013, p. 59). Definitionally and functionally, one's conception of both a corporation and CG depends on one's theoretical perspective. As such, this Section will move from the broadly theoretical, looking at the centrality of both agency theory and stakeholder theory, before moving onto discussing various CG approaches and applications.

2.2.1 Dominant Theories of Corporate Governance

Within the CG literature, Jensen and Meckling's (2000) agency theoretic principal-agent model is seen as the 'traditional' approach (Bertoni et al., 2013, p. 370), even being characterised as synonymous with corporate governance (Lubatkin, 2007, p. 59). A strong literature has demonstrated this, with quantitative studies used to demonstrate that both Western corporate governance models (Viader & Espina, 2014) and U.S. corporate governance norms are substantially dictated by agency theory (Zaman et al., 2022, p. 694).

Agency theory conceptualises firms as a concatenation of contracts, with the agency relationship describing how a party delegates power to another. It operates on the assumption that individuals act with self-interest, prototypically resulting in conflicts of interest between principle-agent goals and information asymmetries (Jensen & Meckling, 2000). Thereby, agency theory typically conceptualises CG's role as being about value-protection, through protecting investors from unaligned management (Bertoni et al., 2013, p. 367) and reducing agency costs to benefit shareholders' fiscal interests (Lund & Pollman, 2021, p. 2574). Despite agency theory's normative position in the management literature, there exists a sizeable literature criticising this theory, highlighting how it maintains an incomplete picture of firms in their relationships with the wider world (Band, 1992; Raelin & Bondy, 2013).

Stakeholder theory instead offers a value-creation perspective of CG, focused on maximising value for all stakeholders whilst avoiding compromise. Conceiving of businesses as the set of relationships between groups who hold a stake in a business's activities, the theory holds that "[n]o stakeholder stands alone in the process of value creation" (Freeman et al., 2010, pp. 24, 27). Due to the nebulous conceptual breadth that the concept of "stakeholder" maintains (Phillips et al., 2003), the concept is used by "various authors in very different ways" (Donaldson & Preston, 1995, p. 66). This has led some scholars to argue for the impossibility of formulating a unified stakeholder theory (Friedman & Miles, 2002) and others to conceive of stakeholder theory as a "genre of theories" (Miles, 2017). The application of stakeholder theory within corporate governance has been extensive and comprehensive (Donaldson & Preston, 1995, p. 71), with the descriptive scope of the theory well documented (Quintelier & Vock, 2022) as well as the benefits of building CG structures that are built using stakeholder theoretic perspectives (Alpaslan, et al., 2009).

Marrying these perspectives, this research will define a 'corporation' as the emergent persona of a nexus of contracts, that can "draw on resources from a variety of different groups" for the end of "long-term value creation" of all stakeholders (Monks & Minow, 2011, p. 47). Corporate governance (CG) will be defined as the collection of ownership and governance structures, both internal and external (Bertoni et al., 2013), that represent the rules and processes that define how decisions are made within a company (Zaman et al. 2022) for the benefit of shareholders and stakeholders.

2.2.2 The Fundamentals of Corporate Governance

An impressive literature has highlighted the essentiality of effective CG. Notable research has found positive correlations between strong CG structures and financial performance (Bruno & Claessens, 2010) as well as between CG and operational performance in advanced economies (Bhagat & Bolton, 2008). Furthermore, extensive metrics and indices of CG performance have been developed (Gompers et al., 2003; Bebchuk et al., 2009). The literature also documents the extensive corporate governance mechanisms available to companies and the significant variety that emerges in the implementation of these tools.

Myriad CG tools exist, each analysed and evaluated in the CG literature, but the most recent decade has seen the literature focus on board committees, executive compensation, financial reporting and disclosure and effective auditing (Wright et al., 2013). Most relevant here, however,

is the literature which has covered how internal governance mechanisms can check on the powers of executive management, with the company board arguably representing both the most predominant CG measure and effective means of control (Naciti et al., 2022, p. 66). The management literature has highlighted the varying duties of a board, all of which depend upon them working alongside the CEO (Tricker, 2020). Within the aforementioned 'nexus', the board is conceptualised as the "corporate nexus" that oversees corporate activities, thus representing a constituent part of corporate governance (Bainbridge, 2008, p. 24).

Among other boards increasing their concern for wider stakeholders when viewing from a stakeholder theoretic perspective, (Wang & Dewhirst, 1992), how board characteristics can affect CSR (Zaman et al., 2022) and improve competitive advantages (Naciti et al., 2022, p. 66), and how board expertise can increase fiscal competency and ethical financial conduct (Velte, 2023). Relevant also is the literature on how CG aids the prioritisation of safety within corporations, with the field witnessing near unified scholarly convergence (Nichols & Walters, 2017). The management literature includes considerable documentation of instances where poor corporate governance practices have comprised safety and had disastrous impacts (Osofsky, 2011; Vaughan, 1990; Abdeldayem et al., 2023). Whilst the transferability of this literature to the AI safety field is strong, this literature has yet to offer a direct contribution.

2.2.3 Corporate Social Responsibility (CSR)

Whilst not an area of direct study within this research, the CSR literature introduces themes and concepts indispensable to studying CG and AI safety, with the critiques, advantages and discourses surrounding CSR and AI safety closely mirroring each other. CSR is predominantly conceptualized in the management literature as a function of CG (Zaman et al., 2022, p. 692) and refers to corporations' voluntary ethical, sustainable, and economic social commitment to stakeholders (Sarkar & Searcy, 2016, p. 1423). CSR is firmly grounded in stakeholder theory and has witnessed extensive attention in the literature in recent years (Tricker, 2020).

Scholars have long highlighted the tensions in CSR between its abstract concept and its actualized implementation (Jones, 1980, p. 59), as well as attendant concerns about the difference between legitimately socially responsible behaviour and corporate image cultivation (Moir, 2001).

Additionally, since the introduction of Fridman's (2007) critiques of CSR, the literature has seen further critiques that a business's sole function ought to be the maximisation of profits (Ibid.) and that the nature of a firm's relationship with broader society beyond a range of primary stakeholders remains an open question (Barnett, 2019). Such critiques are, however, met with an impressive empirical body of research highlighting the efficacy of CSR measures in improving equality and diversity (Hart, 2010) as well as on global issues such as climate change (Eleftheriadis & Anagnostopoulou, 2015).

2.3 Legitimacy and Legitimation

2.3.1 Introducing Legitimacy

Extensively utilised within philosophy, politics (van Leeuwen & Wodak, 1999) and sociology (Glozer et al., 2019), the concept of "legitimacy" has also seen significant application within management organizational studies (Suddaby et al., 2017). Legitimacy will be defined here as the ways that institutions are "explained" and subsequently justified as being "desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" (Berger & Luckmann, 1967, p. 82; Suchman, 1995, p. 574). An extensive management literature has documented the necessity of legitimacy within corporations linked to the 'liability of newness' (Aldrich & Fiol, 1994), which theorises that newer organisations in newer industries have a higher propensity to die (Singh et al., 1986). There also exists some research linking CG and legitimacy in the management literature, with Palazzo & Scherer (2006) advocating to shift the basis of CSR to a morality-based, discourse-derived legitimacy, and Du & Vieria (2012) arguing that legitimacy can be derived from CSR.

Legitimacy acquisition strategies predominantly occupy a managerial perspective and have received extant coverage in the literature, with an interesting facet being how different companies in different industries face differing legitimacy-related challenges. The concept of the 'liability of newness' grounds much of this literature (Aldrich & Fiol, 1994), with an impressive body of work highlighting that newer companies, especially in newer industries, have a higher propensity to die because of external legitimation issues (Singh et al., 1986). As aligned with the differing conceptions of legitimacy, differing solutions are proposed, with Zimmerman & Zeitz (2002), advocating for strategies of conformance, environment selection, environment manipulation, and environment creation, Deephouse et al. (2017) highlight an approach grounded in 'institutionally

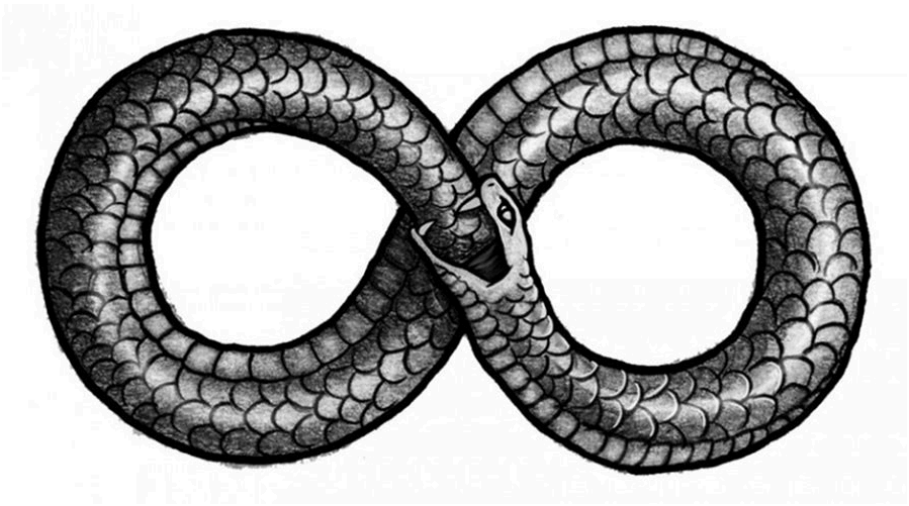
innovating', through changing regulations and performance criteria, and creating cultural-cognitive legitimacy. Aldrich & Fiol (1994), however, advocate for using symbolic language to create trust in new activities whereas Reast et al.'s (2013) study of the UK casino gambling industry found companies pursuing bargaining, capturing, earning, or construing strategies.

2.3.2 Legitimation Differentiated from Legitimacy

However, the focus of this research is on a different form of strategic legitimacy: the discursive acquisition of legitimacy, known as *legitimation*. Whilst closely related to legitimacy, legitimation instead locates legitimacy as an active and continuous process, with language thus representing a central through which social reality is regulated and dictated (van Leeuwen, 2007). It is thus "a process" rather than an object, emerging from actions and interactions (Vaara & Tienari, 2008, p. 987). Conceptualised in this manner, all organisations, actors, groups, and corporations must continually reassert their legitimacy, thus underlining the centrality of legitimation.

Where 'legitimacy' may invoke the image of a legitimate actor resting atop their mountain of accumulated legitimacy, 'legitimation' instead conjures the image of Ouroboros, the snake eternally consuming its own tail. Corporations and institutions must create legitimacy to function, operate and grow, the products of which they must consume and reinvest to continually create and reify that legitimacy.

Figure 1: Ouroboros (IMGBIN, 2019), or The Endless Endeavor of Legitimation



Discursive devices are used to shape, regulate, and manipulate social reality. Such devices include narrative construction (Tienari et al., 2003), cultural toolkits, genre repertoires (Vaara & Tienari, 2008), reframing and rationalisation (Vaara et al., 2003). Despite the management literature still being dominated by quantitative research, analysis of the discursive aspects of legitimacy (Vaara & Tienari, 2008) now represents an increasingly impressive body of scholarly work. This includes work on how top managers use discursive practices to legitimize change within organisations (Demers et al., 2003), discursive legitimation strategies in MNCs (Vaara & Tienari, 2011), and corporate actors' use of discursive devices when seeking to legitimize organizational structuring (Suddaby & Greenwood, 2005). Furthermore, work by (Morsing & Schultz, 2006, p. 333) has highlighted how effective discursive practices around CSR have created corporate legitimacy, with legitimacy management "rest[ing] heavily on communication" (Suchman, 1995, p. 586).

2.3.3 Legitimation Theory

Legitimation theory, principally developed by van Leeuwen, is thus a systematised framework for understanding the language, application and outcomes of legitimation. Drawing on a Weberian typology which understands language as the most vital vehicle through which systems establish legitimacy (van Leeuwen, 2007; Berger & Luckmann, 1967), legitimation theory is grounded in critical linguistics (Fairclough, 2010; van Dijk, 1988) but is focused instead on sociological rather than linguistic categories, demarcating four main methods of legitimation: authorisation, moral evaluation, rationalisation, and mythopoesis (van Leeuwen, 2007).

Within the management literature, legitimation theory has been applied to longitudinal multiple-case studies, in the analysis of new venture legitimation (Turcan, 2012), as well as for understanding legitimacy issues within sustainable business model innovation (Biloslavo et al., 2020). There exists, at present, limited research that specifically utilises legitimation theory as its primary theoretical lens, but there are significant studies focusing on discursive legitimation that reference legitimation theory (Vaara & Tienari, 2008). Such studies include Luyckx & Jansenns' (2016) work analysing the historical dimension of discursive legitimation within controversial multi-national corporations over time, as well as Beelitz and Merkl-Davies' (2011) analysis of how 'CEO-speak' is utilised to restore organisational legitimacy in the wake of crises, which found that CEOs "use discourse to manufacture organisational audiences' consent" (Beelitz & Merkl-Davies, 2011, p. 115). Further research has looked at the discursive legitimation of corporations in the wake

of perceived scandals and alleged wrongdoings (Breeze, 2012), as well as Heinzmann & Fox's (2019) critical discursive analysis of HR managers' struggle for legitimacy during processes of change.

Finally, the review of the literature found no comprehensive work produced on AI safety and legitimation. The Oxford Handbook on AI Ethics does refer to legitimacy multiple times, highlighting briefly that stakeholder interaction increases legitimacy (Dubber et al., 2020, pp. 89), but it makes no contribution to discursive strategy, no reference to legitimation, and includes no meaningful discussion about corporate governance and legitimacy. Thereby, this represents a significant shortcoming of the literature that ought to be addressed.

2.4 Concluding the Literature Review

To briefly conclude this review, the AI safety literature was highlighted as being rich from a technical, regulatory and legal perspective, but the management literature's contribution has thus far been disappointing. The literature on corporate governance was found to be replete with compelling, intelligent and illuminating research, but again failed to contribute meaningfully to AI safety. Finally, whilst legitimation research has gained impressive contributions in recent decades, with ample research with applicability to corporate governance and certain facets of AI safety, at present there exists no literature making a direct link between the two.

Chapter 3: Methodology

This paper's research aims of this paper are intrinsically interlinked with the methodological approach applied, which are as follows:

> RQ1: What are the textual and intertextual legitimation techniques that leading AI labs use to legitimate their novel corporate governance structures?
>
> RQ2: How do leading AI labs conceptualise their responsibilities in relation to themselves and corporate governance?
>
> RQ3: What wider implications do these findings have for AI safety?

Central to the methodological approach utilised here is the synthesis of two CDA frameworks, which form a continuous thread throughout this chapter. In unifying van Leeuwen's (2008) and Fairclough's (2010) CDA methodologies, significant care has been taken to ensure their vital tenets and respective advantages are retained.

3.1 Research Philosophy

My ontological stance is best described as critical realism, with the realist component denoting that 'the world' exists regardless of our knowledge of it and the critical component highlighting the social world's difference from the natural world's because the former is "contingent upon human action" (Fairclough, 2010, p. 4). The nuance here is grounded in the distinction between *construal* and *construction*, as discursively construing the world subjectively does not mean the natural world is thus constructed by our discourses.

Regarding the epistemological perspective, I maintain a critical interpretivist perspective. Knowledge is understood as describing all the contents of consciousness, as well as "all kinds of meanings used […] to shape the surrounding reality" (Jäger, 2011, p. 32). Knowledge production is conceived of being received and mediated through discourses, power relations and cultural contexts, with the interpretivist slant derived from an emphasis on the contextual, dialectical and reflexive nature of agents' interactions with structures (Grix, 2002, p. 183). In turn, the critical

element describes the necessity to analyse power relations and to employ a rigorously analytical approach to understand how knowledge is produced and legitimated.

This philosophy is uniquely well-disposed for addressing this paper's central concerns, of corporate governance, legitimacy and AI ethics, as the critical realist perspective enables consideration of both the material reality of AI systems and the social constructions surrounding their development. As outlined in Section [2.1], misaligned or misused AIs have the potential to restrict human freedoms and a discursive consideration of the natural world's construal is essential in preventing this. Epistemologically, the contextual component allows for the recognition that governance practices are shaped by specific cultural norms, practices, and organisational contexts. The dialectical factor acknowledges the relationship between structures and agents in the formulation of AI policies, whilst the reflexive component underlines the need for the constant questioning of this author's own biases and assumptions, essential within the rapidly evolving field of AI.

3.2 Design of the Study

Regarding overall strategy and approach, the research conducted here is a documentary study. As argued by Vaara and Tienari, (2008, p. 987) "[f]rom a discursive perspective, the starting point for any analysis of legitimation is the notion that senses of legitimacy are created in relation to specific discourses". As such, through analysing the chosen corporate documents, detailed in Section [3.3], this approach will glean valuable insights into the companies' official and unofficial policies. This strategy also enables the tracking of different discourses across different organizational contexts, which is indispensable here due to the differing companies analyzed.

In analyzing corporations' legitimation strategies, this was done using naturally occurring data. Through analysing publicly available data in the form of publications and podcasts, it was possible to gain insight into how these companies and their leaders present their corporate governance structures (Golato, 2017). The benefits here are two-fold: firstly, this approach allows for the observation of unprompted legitimation strategies, thus further limiting the influence of my own biases in the research; and secondly, the publicly available data published enables accessibility to these companies' perspectives. It is important to reiterate, however, that whilst this data is 'natural', it is certainly not neutral. By conceptualising discourse as "a socially constructed knowledge of some social practice" formulated in specific social contexts (van Leeuwen, 2008, p. 6), language is seen as inextricable from both reality and social practice.

Central also to this research is the critical stance that has been adopted. As aligned with my research philosophy, this stance operates on the belief that corporate legitimation strategies surrounding AI labs' corporate governance policies are not neutral artefacts but instead complex constructions that are both embedded in and shaping a broader range of socio-political contexts. Discourse is "relational, it is dialectical, and it is transdisciplinary" (Fairclough, 2010, p. 3), so rather than taking AI companies' published documents and their CEO's musings at face value, this methodology will seek to question the underlying assumptions and intentions, through critically examining the language and framing of their corporate communications, and highlight not only what is said, but also what is implied.

3.3 Data Collection

A process of judgement-guided purposive sampling was used for data selection (Saunders et al., 2023, p. 322), to enable the selection of texts that included extensive discussion of AI labs' corporate governance structures. From a methodological perspective, this significant freedom was enabled by there being no stipulations "concerning data collection requirements" in the works of Fairclough and van Leeuwen (Meyer, 2011, p. 24). Four texts were selected representing two AI labs [See: Table 1]. Regarding company selection, Anthropic and OpenAI are, alongside Google DeepMind and Meta AI, two of the world's leading companies in the race toward AGI. Different from the latter two, however, they maintain more novel CG structures and have engaged in more significant discursive legitimation.

Table 1: Selection of Texts for Analysis

Text	***Introduction to Text, and Location***
Text 1: OpenAI (2024) "Our Structure"	*OpenAI's introduction to and legitimation of their corporate governance structure, published on their official website* *Full coded text is located in Appendix 1*
Text 2: Anthropic (2023) "The Long-Term Benefit Trust"	*Anthropic's introduction to and legitimation of their corporate governance structure, published on their official website* *Full coded text is located in Appendix 2*

Text 3: Altman (Fridman, 2024) "Sam Altman"	*Excerpt from an interview with OpenAI's CEO, Sam Altman, on the Lex Fridman podcast* *Full coded text is located in Appendix 3*
Text 4: Amodei (Patel, 2023) "Dario Amodei"	*Excerpt from an interview with Anthropic's CEO, Dario Amodei, on the Dwarkesh podcast* *Full coded text is located in Appendix 4*

These texts were selected for several reasons. Firstly, by selecting two different forms of text, the intertextual analysis was enriched by allowing for greater discussion and comparison. Secondly, in selecting texts that broadly mirrored each other in both companies, an increasingly accurate comparison between the two companies was enabled. Thirdly, through selecting texts that feature both the official written publication of an AI lab, as well as the opinions shared by the respective CEOs in a more informal setting, a comparative analysis of the varying legitimation techniques employed within each company was enabled. Finally, it allowed also for a comparative analysis of the CEO and the official company line.

Each document's original publication was in the researcher's native English, thus negating any methodological issues associated with corrupted translation (Saunders et al., 2023, p. 159), and all four texts were published between August 2023 and March 2024. This timeframe was selected as it was sufficiently narrow to enable fair comparisons and sufficiently recent to enable relevant research. To ensure the accurate representation of the CEO's opinions, I cross-checked each podcast and transcript, with the nature of Texts 1 and 2 not necessitating such measures.

A line-numbered and coded copy of each text can be found in Appendices 1, 2, 3, and 4. Throughout the analysis and discussion [Chapter 4], references to "Text" in the summary tables and throughout refer to these Appendices.

3.4 Data Analysis

3.4.1 CDA Design

A critical discourse analysis (CDA) was deemed the most effective data analytical approach for addressing the central research questions. Discourse analyses endeavour to analyse the knowledge

created in discourses, as well as the antecedents of their presupposition (Jäger, 2011, p. 33). CDA maintains this same goal, but also deploys a *critical* social analysis, to combine a "critique of discourse" with an "explanation of how discourse figures in existing social reality" (Fairclough, 2017, p. 35). Whilst word frequency analyses can provide meaningful insights, they have not been utilised here, as statistical numerical representation can fail to account for the meaning and power assertions that lie within texts (van Leeuwen, 2013, p. 42).

Fairclough, having traditionally focused upon linguistic manifestations "of dominance, difference and resistance" in discourses of Marxist social conflict (Meyer, 2011, p. 22) introduced a three-dimensional model, which consists of social practice, discursive analysis, and linguistic analysis. The linguistic facet focuses on the close linguistic analysis of texts, with techniques including "phonology, grammar [and] semantics" (Fairclough, 1992, p. 194), whilst the discursive analysis focuses on texts' intertextuality, acknowledging that texts necessarily exist within preexisting 'orders of discourse' (Fairclough, 1992, p. 192). The analysis of the wider social context then enables the proper contextualisation and comprehension of the production of a certain discourse, by enabling the researcher to situate the text in its certain historical, cultural, and economic contexts (Fairclough, 2010, p. 3). However, as Fairclough (1992, p. 212) highlights, "linguistics is still dominated by a formalism which has little time for integrating linguistic analysis into interdisciplinary frameworks". Thereby, an issue of accessibility emerges for non-linguists when conducting the textual analysis stage of Fairclough's methodology. As such, van Leeuwen's legitimation framework will be integrated into this stage to solve this issue.

Most pertinent to this paper is van Leeuwen's theoretical and methodological work on legitimation. van Leeuwen (2008, p. 105) posits that "[l]anguage is without doubt the most important vehicle" through which institutions assert their legitimacy, meaning language and legitimacy are inextricably bound to one another. His method of CDA is constituted of a close-point textual analysis that enables the discerning of how language is used to legitimate and delegitimate various actors. The accessibility issue deriving from Fairclough's CDA approach can be resolved by utilising Systemic-Functional Linguistics (Fairclough, 1992, p. 212), which van Leeuwen's (2007) legitimation theoretic CDA approach is grounded in. Through providing categories and manifestations of legitimation, van Leeuwen's [See *Table 2* for full descriptions] textual analytic framework enables both deeper integration of legitimation theory and accessibility for non-linguists (van Leeuwen, 2008, pp. 105-106). Whilst van Leeuwen's (2013) 'social actor theory' is not applied here, his language of 'actors' will be utilised throughout.

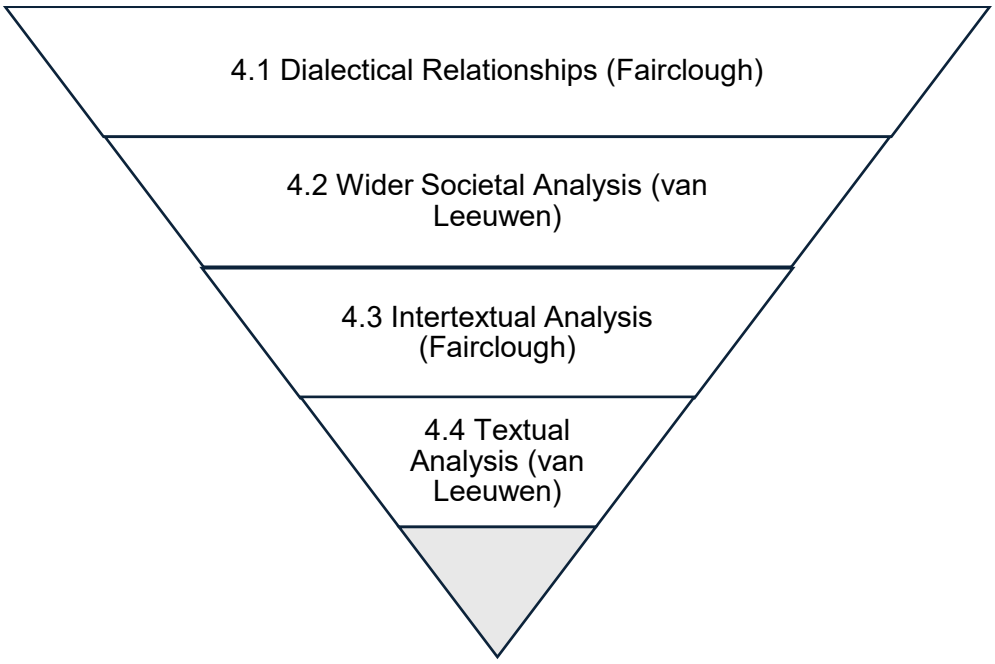
Therefore, to gain the benefits of both CDA approaches, van Leeuwen's textual framework will be integrated into Fairclough's framework. There exists a strong precedent in legitimation

studies to apply two different CDA approaches, such as with van Leeuwen & Wodak's (1999) integration of a discourse-historical approach and an analytical approach. However, precedent was not found in the existing literature for the integration of Fairclough's and van Leeuwen's CDA frameworks. In the analysis [Section 4], the synthesis of the frameworks can be understood as follows: Fairclough's CDA methodology informed the macro-level (social practice) and meso-level (discursive and intertextual) analyses, whereas van Leeuwen's work on legitimation was used to sustain the micro-level analyses in place of Fairclough's linguistic analysis. A visual depiction of this can be found in *Figure 2*.

3.4.2 CDA Implementation

This Section focuses on demarcating the various stages of analysis, as well as signposting where they can be found in this dissertation. A visual depiction of this Section can be found in *Figure 2*, along with an attribution of the theoretical grounding.

Figure 2: Visual representation of the Stages of Analysis, based on Fairclough's (2010) and van Leeuwen's (2007) CDA methodologies



The first part of the analysis involved illuminating the dialectical relationships between social practices and the texts using Fairclough's (2010, p. 3) methodology, including how they were produced and how the forms of discourse analysed construct certain forms of reality. For this study that involved exploring how podcasts construe social reality and how corporate websites affect consumers' understanding and perceptions of company practices, by consulting both the management literature and wider contemporary research [Section 4.1].

The second phase of the analysis again utilised Fairclough's approach and involved evaluating the companies' broader corporate structures and power relations to understand how they shape, and are shaped by, these discourses. This was achieved through analysis of the facts posited in the texts, which were then cross-checked with multiple other sources for truth and validity, as well as analysing a wide breadth of contemporary reports. The results of this can be found in Section [4.2], structured under three dominant themes and provide increased context for the legitimative strategies.

The third phase of analysis, also using Fairclough's methodology, was the intertextual analysis, which focused on the 'orders of discourse', which are constituted of the language that is used, the traditional tropes that are drawn upon, and the predominant conventions of communication (Fairclough, p. 242, 2010). The necessity of this phase is grounded in the belief that discourses do not emerge in isolation of each other, instead drawing on discursive narratives, with this phase of the analysis thereby essential in "linking text to context" (Fairclough, 1992, p. 213)

The fourth stage involved the textual analysis, using van Leeuwen's (2007) legitimation CDA. Having collected the data, a process of initial reading was undertaken, through which preliminary patterns, themes, and observations were identified. Following this came the development of a theory-driven deductive coding scheme, predominantly based upon van Leeuwen's categories of legitimation [See Table 2]. Having developed the coding scheme, a process of iterative manual coding was deployed across all four documents to ensure the authenticity of both voice and interpretation. Having coded for linguistic features and identified discursive features, the attention of research then turned to a thematic analysis, a methodology not tied to any research approach (Saunders et al., 2023, p. 665), through which dominant themes and patterns were identified which subsequently formed the structuring of Sections [4.4]

Table 2: Summary of van Leeuwen's Categories of Legitimation, tabulated by the author, using van Leeuwen (2007)

Form of Legitimation	*Definitions*
Authorisation	**Legitimacy derived through reference to some other norm, actor, or institution, that is deemed legitimate**
Personal authority	Owing to role or status within an institution
Expertise authority	Owing to expertise rather than status
Role model authority	Opinion leaders, deemed "'wise', 'experienced', 'cool', or 'smart'"
Impersonal authority	Laws, policies, regulations, and commandments
Authority of tradition	Habit, tradition, practices, and custom
Authority of conformity	Owing to the authority of the majority and popular consensus/action
Moral evaluation	**Legitimacy based upon values and ethics, without further justification**
Evaluation	Evaluated adjectives like "good" and "bad", but also indirect words, such as "healthy", "normal", and "useful"
Abstraction	The abstracting of practices in a way that moralises them
Analogies	Denotes something as good through equating it with an alternate activity associated with positive values
Rationalisation	**Legitimacy derived from purposes, always with a moral slant; two categories of instrumental (IR) and theoretical (TR) rationalisation**
IR: Goal orientation	"I do *x* in order to do (or be or have) *y"*; can be either implicit or explicit
IR: Means orientation	"I achieve doing (or being or having) *y* by *x*-ing", or "*x*-ing serves to achieve being (or doing or having) *y*"
TR: Definitions	Use of definitions, to ground the truth of what is said
TR: Explanation	Asserting something to be appropriate in the presence of certain social actors
TR: Predictions	Assessing the way things will be based upon current assumptions
TR: Scientific	"Differentiated bodies of knowledge" designed to legitimise institutions
Mythopoesis	**Use of storytelling to legitimate various actors, or delegitimate others**

Moral tales	Stories in which protagonists who conform to legitimate social practice or restore social order are rewarded
Cautionary tales	Stories that highlight the negative results of non-conformance to norms
Single Determination	Narrative that offers a straight-forward justification or explanation
Over-Determination	Narrative that offers multiple, often conflicting, explanations

3.5 Approach to Theorising

The research presented here involves both deductive and abductive elements. A critical realist ontological foundation is typically associated with abductive reasoning, as it greater allows for a continuing transition and complexified interplay between data and theory (Saunders et al., 2023, p. 159). As highlighted in Section 2.4, despite drawing on extensive work conducted from several standpoints within the management literature, the novelty of AI safety and its' lacking coverage within the field necessitates some abductive elements. Points of discussion and analysis relating to existing frameworks will be drawn upon and theorized deductively, but an abductive approach will be utilized in novel areas, to enable greater academic creativity.

3.6 Ethical Considerations and Limitations

Due to the utilisation of secondary publicly available data, minimal ethical considerations were required. Clearance from an impartial ethics body was obtained, and the author took extensive care to ensure the proper representation of the sources.

The central concern within CDA stems from the researcher's subjective interpretation of texts and these biases working their way into the analyses (Saunders et al., 2023, p. 694). Some have argued that CDA ought thus not to be deemed analysis but instead "biased interpretation" (Widdowson, 1995, p. 169), although this claim is countered by Fairclough's argument that the undetermined, openly possible results of CDA do indeed make it analysis (Meyer, 2011, p. 16).

Whilst I believe it is impossible to perform any research free from *a priori* judgements (Meyer, 2011, p. 17), I believe also that CDAs are more liable for biased interpretation. The primary means for mitigating biases stem from a constantly reflexive process, achieved by routinely questioning assumptions, judgements, and interpretations.

Other measures implemented centred upon maximal transparency, including providing clear rationales for textual selection [3.3] as well as detailed accounts of analytical methods and their rationale [3.4] (Saunders et al., 2023, p. 729). As a further measure, it was thought beneficial to briefly outline my ideological grounding on issues relevant to this research. My perspective about the development of AI and its ensuing effects on humanity is best described as more pessimistic than optimistic. Politically, I maintain a classical liberal viewpoint and ethically I think in primarily consequentialist terms.

Further limitations of this study include limited generalisability, owing to the focus on two companies and the depths of the analysis. However, the implementation of the contextual and discursive analyses aids in facilitating the transferability of the conclusions (Fairclough, 1992)

3.7 Summary

This chapter has served to underline the core methodological decisions made. Through highlighting the key tenets and benefits of Fairclough and van Leeuwen's approaches to CDA [Sections 3.4.2 and 3.4.4], an extensive account of their integration [Figure 2] and implementation [Section 3.4.2] was enabled. In detailing my research philosophies [3.1] and general study design [3.2], solutions to the various limitations of this approach [3.6] were enabled.

Chapter 4: Findings, Analysis, and Discussion

The CDA undertaken illuminated a comprehensive, consistent, and persistent use of a variety of legitimation techniques by all actors. The findings focus on the forms of legitimation techniques (van Leeuwen, 2007) utilised by the various actors (RQ1); the analysis centres on how the various social actors both legitimating *themselves and their actions* as well as legitimating *their corporate governance structures* (RQ2)*;* and the discussion component will focus on placing this research in the wider literature, as well as understanding how legitimation links to corporate governance and AI safety (RQ3).

The following chapter displays the findings of the CDA, the analysis of the data, and the discussion interpolated throughout. The chapter's structure follows the same structure as outlined in Figure 2. Within each section, several themes were identified, around which the findings, analysis and discussion are developed. Definitions for each legitimation technique identified can be found above in Table 2, which will be an indispensable tool for the comprehension of this chapter.

4.1 Dialectical Relationship Analysis

The dialectical analysis addressed two forms of discourse: podcasts, and official corporate documents. From a critical perspective, both podcasts and official corporate documents implicate increased trust, thus representing indispensable tools for corporations' legitimation in the 21st century.

Since the advent of the internet, companies have used websites to communicate their values, beliefs, and competitive advantages to companies (Cox et al., 2008) and MNCs now almost ubiquitously exhibit corporate websites (Tang et al., 2015). An analysis of public perceptions of corporate press releases found that when companies claim that their official announcement is true on their publication channels, readers perceive greater organisational transparency (Kim et al., 2014, p. 811). The management literature has drawn links between legitimacy and non-financial disclosures, with Morsing & Schultz (2006, p. 333) underlining they are "produced to inform and convince public audiences about corporate legitimacy". Evidence of OpenAI's [Text 1] and Anthropic's [Text 2] awareness of this was found primarily in their intertextual strategies, through

their reserved and legitimating uses of militaristic, scientific and governmental orders of discourse [See Section 4.3].

Regarding the dialectical relationship between podcasts and construed reality, podcasts have radically altered how people consume information and derive entertainment. Recent reporting has highlighted podcasts' unique influence, with podcasters being twice as influential in opinion-changing than social media due to greater perceptions of trust and intimacy (Magma, 2023). This research is supported by the academic literature, with Jarrett (2009) highlighting how podcasts uniquely amplify private talk into the public sphere and Waddingham et al. (2020) highlighting how corporate podcast appearances can increase listeners' understanding of an organisation. As such, they offer podcast guests the opportunity to communicate their values and aspirations in an intimate setting and offer CEOs an opportunity to further create a narrative and legitimate themselves and their actions. Textual evidence for this familial dialectical relationship, as well as AI corporation's CEOs' probable perception of it, was found in their extensive mythopoeic narrative legitimation [4.4.3].

As such, the nature of these dialectical relationships results in these discourses representing indispensable vehicles for corporations and their CEOs to construct narratives (Vaara & Tienari, 2011), manipulate public perceptions, and ultimately legitimate themselves and their actions.

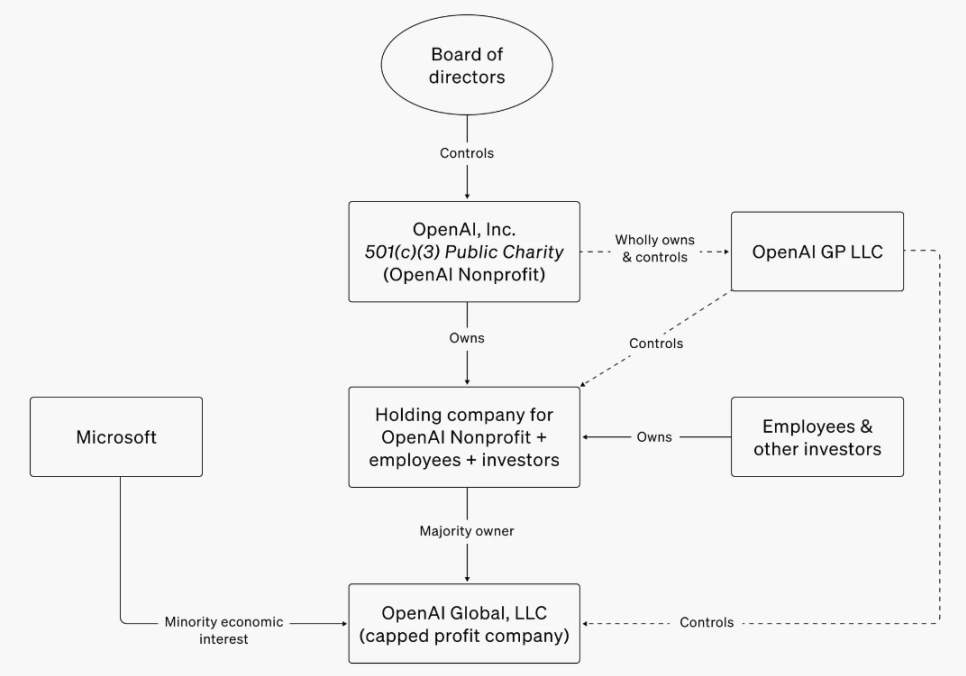
4.2 Contextual (Socio-Corporate) Analysis

The contextual analysis, conducted as outlined in Section [3.4.2], illuminated interesting patterns around legitimation, AI safety, and corporate governance, which provided further context and grounding for the textual [4.4] and intertextual analyses [4.3].

Firstly, OpenAI and Anthropic have formulated markedly different corporate governance structures. OpenAI was founded in 2015 as a non-profit, but the corporate structuring was revised in 2019 to have OpenAI Inc. (a non-profit organization) owning a capped-profit company (OpenAI LLC) [See Figure 3] that allows 100x returns on initial investments (Bansal, 2024). This mixed-profit structuring has been highlighted in the academic literature as "highly unusual for cutting-edge companies" (Tallarita, 2023) and has been criticised for "lack[ing] legal standing and recognition" (Andhov, 2024, p. 11). Whilst this structure enshrines in its CG structure the non-profit controlling the for-profit, the blurred lines of responsibility between the branches serve to veil rather than elucidate where power in this structure truly resides. As can be seen in comparing

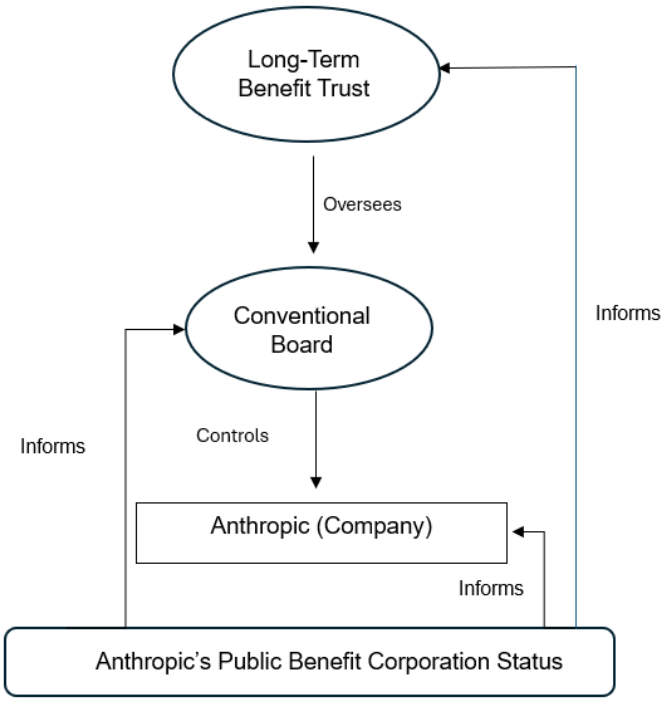
Figures 3 and 4, OpenAI's structuring is far more complicated, whilst failing to highlight how this complexity may improve AI safety.

Figure 3: Visual representation of OpenAI's corporate governance structure (Figure taken from (OpenAI, 2024))



Anthropic, alternatively, incorporated as a Public Benefit Corporation (PBC), thus legally enshrining their commitment to stakeholders and the wider public and formulated a new form of corporate governance, dubbed "The Long-Term Benefit Trust" (Anthropic, 2023). PBCs represent a novel legal form of corporation (Kurland, 2017) and have been argued in the CG literature to represent an "ethical step toward empowering socially committed commercial entities" (Hiller, 2013), although others have highlighted concerns surrounding actionable accountability (Cummings, 2012). In the legitimacy literature, Segal (2017) has argued that PBCs represent a meaningful step towards granting corporations greater legitimacy in the 21st century. Anthropic's LTBT essentially enshrines a board *above* the board, who maintain a new class of stock which enables them to appoint a majority of Anthropic's board members [See Figure 4]. Locating Anthropic's structure within the management and legitimacy literature, their LTBT may represent an actualisation of Marchant & Wallance's (2013) proposition for meta-level oversight structures that govern the governance of emerging technologies.

Figure 4: Anthropic's Corporate Governance structure (Figure created by the author, based upon (Anthropic, 2023))



Such developments ought not be rejected ought of hand, as AI companies cannot "rely on traditional corporate governance to protect the social good" (2023, p. 1). However, the novel nature of these CG structures means they lack inherited legitimacy, thus increasing these actors' necessity for discursive legitimation.

The second theme that emerged from the contextual analysis was that both OpenAI and Anthropic have received significant influxes of capital. OpenAI had received $130 million of grant money by 2019 from its founding donors, but since shifting in that same year to the mixed-profit model, they have received almost $13 billion in investment from Microsoft and billions more from independent venture capitalists (The Guardian, 2024). Microsoft's extensive investment has given them proprietary rights to new OpenAI technologies, a 49% share of the business, a non-voting position on the board, and the rights to a profit share (Bansal, 2024).

Turning to Anthropic, after receiving a $500 million investment from the then-reputable FTX (Knauth, 2024) they have since received investment from Amazon totalling $4 billion, in

return for a foothold in the 'AGI race', access to Anthropic's leading models, and an unpublished equity stake. Such capital influxes from these corporate giants increase AI corporations' power and capacities, thus providing new avenues to legitimate themselves [Section 4.4.1] and greater incentives for prioritising profit over AI safety.

In predicting the future of corporate governance, Tricker (2020, p. 36) argued that "new forms of corporate identity will emerge" and that "new sources of capital will replace existing markets". However, rather than treating these as non-causal events, it may be more apt to conceptualise their interrelation. OpenAI's and Anthropic's corporate governance structures have not been born in isolation or formulated solely to benefit humanity [Section 4.4.2], but also to cater to the investments that these companies have accepted. As Talaria (2023, p. 3) has highlighted, "Even creative governance structures will struggle to tame the profit motive [within AI labs]", a concern which can only be exacerbated by these significant influxes of capital.

The third theme that the contextual analysis surfaced was the differing stability and instability of each company's corporate governance structures. Anthropic's corporate governance structuring has been devoid of notable scandal and instability, which is reflected in their academic and intertextual orders of discourse [Section 4.3.2]. OpenAI's corporate governing has been more tumultuous, however. In November 2023, the board sacked the CEO, Sam Altman, alleging he had not been sufficiently open and candid with the board. Within three days, however, following a series of demands by Altman, the employees and Microsoft, Altman was reinstated and assumed a voting position on the board, the board members who voted to oust Altman were removed, and Microsoft accepted a non-voting position on the board. Extensive attempts to legitimate this are made by Altman [See Section 4.4.3] but in Text 1, OpenAI neglects to make any mention of the instability.

Discussing this within the bounds of the literature, this creates significant concerns regarding AI safety, CG and legitimacy within OpenAI. As the literature highlights, high degrees of "social organisation" in the form of effective CG enables the kinds of stability and cohesion that technological innovation requires (Lazonick & O'Sullivan, 1996; Tylecotea & Ramirez, 2006, p. 163). It thereby stands to reason that the instability at OpenAI has affected effective innovation, of which their AI safety development is a core part. Evidence for this is furthered by some of the past decades' worst lapses in corporate safety standards being attributed to ineffectual CG functioning (Osofsky, 2011; Vaughan, 1990; Abdeldayem et al., 2023). Whilst some sections of the literature advocate for smaller boards "dominated by insiders" in high-tech companies (Raheja, 2005), the literature also highlights issues with companies' boards maintaining insufficient cognitive distance

(Baysinger & Butler, 2019), especially within AI companies (Tallarita, 2023, p. 6). This may therefore create cause for concern around AI safety that OpenAI's CEO sits on its 'independent' board (OpenAI, 2024), thus further blurring the lines between OpenAI's for-profit and non-profit operations.

4.3 Intertextual Analysis

In the four texts analysed, varied orders of discourse were highlighted. The findings of the analysis have been grouped into three key themes of discourses, to enable clarity as well as evaluative depth.

4.3.1 Militaristic Order of Discourse

The first of these is the extensive utilisation of a militaristic order of discourse. Found in OpenAI's, Altman's and, to a lesser degree, Anthropic's discourses, there exists many further instances than just those listed in Table 3. This order manifested itself primarily through a variety of word choices. Corporate language at large is replete with militaristic analogies, capitalist competition is often analogized to a fight for economic survival, and Sun Tzu's *The Art of War* regularly tops out lists of the most important business books. The use of such language in discourses, however, is not innocuous, serving to shape social reality. The military is predominantly deemed a legitimate institution, due to a perceived necessity of defence, meaning that through drawing on this discourse these actors implicitly legitimize themselves by association with the militaries' underlying "validity claims" (van Leeuwen, 2007, p. 101) and draw also on the impersonal, personal and impersonal authorization that militaries maintain.

Table 3: Militaristic Orders of Discourse

Reference	***Quotes from Text***	***Order of Discourse***
Text 1, Line 57-58	"the for-profit would be tasked with marshalling the resources to achieve this while remaining duty-bound to pursue OpenAI's core mission"	*Militaristic*

Text 1, Line 5	"a chassis for OpenAI's mission"	*Militaristic*
Text 1, Line 32-33	"jeopardizing our mission"	*Militaristic*
Text 3, Line 98-99	"You do that on the battlefield. You don't have time to design a rigorous process then"	*Militaristic*
Text 3, Line 121	"How can we deploy this"	*Militaristic*
Text 2, Line 169	"The Trust structure was "red teamed"	*Militaristic*
Text 2, Line 133	"who helped our leadership design and "red team" this structure"	*Militaristic*
Text 2, Line 11	"our mission"	*Miltaristic*

Analysed through the lens of legitimation theory, extensive legitimation techniques are deployed here to legitimate each of the respective actors and their actions. Primary among these is the implicit means-oriented theoretical rationalization [See 4.4.4] that this order of discourse implies. War has long been conceptualized as a 'necessary evil' within which conventionally unjustifiable means become accepted. The continued evocation of the term "mission", as well as "duty-bound" further reflects this, whilst the coupling with words like "jeopardize" and "endeavour" deepens these links.

In saying, "You do that on the battlefield", Altman morally evaluates his actions by way of analogy and invokes a means-end rationalization of his actions. With each of the other examples in Table 3, the rationalization is more implicit. Language like "red-teamed", a term commonly used in military strategy, furthers this connotation, whilst phrases like "deploy" and "marshalling" imply top-down commandment, thereby implicitly legitimating the leadership of OpenAI using personal and impersonal authority. The invocation of "marshalling" here is used to describe how OpenAI's non-profit interacts with their for-profit company, thus representing a further attempt to legitimate their mixed-profit corporate governance structure.

When considering the implications of this order of discourse in relation to AI safety, there could be cause for concern. As highlighted in the Literature Review [Section 2.1.3], Armstrong et al.'s (2016) Nash equilibrium modelling of an AI arms race predicted negative consequences, especially if enmity between 'teams' exists. Through using a militaristic discourse, OpenAI,

Altman and Anthropic imply that they conceptualise AI's development to be analogous to competition at best, and war at worst. This also further suggests that calls for cooperation and collective action within AI corporations may lack actionability (Askell et al.,2019).

4.3.2 Governmental Order of Discourse

The second thematic order identified is described here as a governmental order of discourse [See Table 4]. Most present in Anthropic's official publication (Text 2), this order of discourse is conceptualised as one that evokes concepts of government, constitutions and politics, and was identified primarily by language choices which evoke such institutions.

Through institutions like free electoral systems, strong constitutions and impartial courts, democratic governments are often deemed in the Western world to be the archetypal representation of a legitimate entity. In framing their corporate governance structures in this language, Anthropic casts themselves as a quasi-governmental figure and subliminally draws on myriad legitimation techniques by doing so. Word choices such as "amended", "supermajorities" and "one year-terms" directly invoke constitutional politics, whilst lines like "carefully balance[ing] durability and flexibility", "checks and balances", and "insulate the Trustees from financial interests" implicitly evokes the U.S. constitution [See Table 4 for references]. Furthermore, underlining that the LTBT represents an "independent body", as well as emphasising their "direct representation" of shareholders, continues the allusions to the core principles of democratic government. By likening themselves to an institution that maintains a certain level of legitimacy they heavily utilize moral evaluation by way of analogy, whilst legitimating themselves also using both personal and impersonal authorization. Parallels can be seen here with van Leeuwen & Wodak's (1999, p. 104) CDA of immigration policies, in which agents treated legal judgements "as an autonomous institution which requires no anchoring in some overarching moral order". In utilising this order of discourse Anthropic take this one step further, by casting themselves as the institution itself.

Table 4: Governmental Order of Discourse

Reference	***Quotes from Text***	***Order of Discourse***
Text 2, Line 97	"amended our corporate chapter"	*Governmental*

Text 2, Line 31	"We have therefore designed a process of amendment that carefully balances durability with flexibility"	*Governmental*
Text 2, Line 144	"sufficiently large supermajorities"	*Governmental*
Text 2, Line 161-162	"Trustees serve one-year terms and future Trustees will be elected by a vote of the Trustees.""	*Governmental*
Text 2, Line	"externalities tend to manifest themselves progressively more, making checks and balances more critical."	*Governmental*
Text 2, Line	"agreements are designed to insulate the Trustees from financial interests"	*Governmental*
Text 2, Line	"independent body"	*Governmental*
Text 2, Line 101	"to ensure that our investors' perspectives will be directly represented"	*Corporate; governmental*

At the surface level, an AI corporation conceptualizing itself as having governmental responsibilities, to all of a nation or even all of humanity, may appear as a positive for AI safety. Considering this from a critical perspective, however, this represents an empty discursive legitimation strategy. As highlighted before, Anthropic has not received assent through a popular vote, nor does their CG structure enshrine such a system and moralizing their CG structure as including an "independent body" is neither a guarantee of independence nor safety. In Text 1 (Line 74-75), OpenAI frame Altman's independence, arguing that he "does not hold equity directly", instead only holding a "small investment" he made before joining OpenAI. This stake is now worth in excess of a billion dollars, thus underlining the manipulative use of discourse to legitimate their CG structures. Furthermore, even if these boards are independent, social responsibility does not follow as a necessary result (Tallarita, 2023, p. 6), with corporate boards often neglecting social responsibility despite maintaining independence (Rashid, 2021). As such, the utilisation of the governmental order of discourse may represent a hollow attempt to legitimate.

4.3.3 Scientific and Academic Orders of Discourse

The third theme highlighted a scientific order of discourse utilised by Amodei [Text 4] and an academic order, used by Anthropic [Text 2].

In his podcast interview, Amodei continually evoked a scientific discourse, conjuring images of both the scientific method and process, as well as likening perceptions of him and his company to the institution of science. In highlighting that he and fellow founders were physicists, as well as using language like "calculus" and "thinking about things intellectually", Amodei legitimates himself using multiple techniques. These include scientific theoretical rationalisation, through which he appeals to science's "differentiated bod[y] of knowledge" (van Leeuwen, 2007, p. 104), as well as expertise authority and the authority of tradition. The implicit rationalisation applied here could be construed as follows:

P1: We are scientists, thereby making us rational, selfless, and well informed

P2: We will bring these qualities to our work in developing CG systems and AI

C: This therefore makes us legitimate actors.

Such an interpretation is given greater weight by Amodei eschewing the prototypical perception of a CEO, instead asking to be seen as "boring and low profile", an image unfortunately often associated with scientists, and viewing his company not in terms of the militaristic, highly competitive vehicle evoked by OpenAI, but instead, like science, as a "nameless, bureaucratic institution". The academic tone highlighted in Anthropic's official publication seeks to serve similar ends but with different implied connotations. Language like "some wonder, therefore" invokes images of a balanced consideration of views following a complete surveyance of a theoretical debate. This is given greater weight by the penultimate quote in Table 5, which performs the same function but with a tone of paternalistic condescension. Again, this enables Anthropic to draw upon personal, traditional and expertise authority, as well as scientific and definitional theoretical rationalization (van Leeuwen, 2007).

Table 5: Scientific and Academic Orders of Discourse

Reference	***Quotes from Text***	***Order of Discourse***
Text 4, Lines 56-57	"it's all part of the calculus"	*Scientific*
Text 4, Line 67-68	"I want to defend my ability to think about things intellectually"	*Scientific; academic*

Text 4, Lines 75-76	"I want people to think in terms of the nameless, bureaucratic institution and its incentives"	*Scientific*
Text 4, Lines 63-64	"if people think of me as being boring and low profile, this is actually what I want"	*Scientific*
Text 4, Lines 40-41	"And because several of our founders, myself, Jarad Kaplan, Sam McCandlish were physicists"	*Scientific*
Text 2, Line 31-33	""At Anthropic, our perspective is that the capacity of corporate governance to produce socially beneficial outcomes depends strongly on non-market externalities"	*Academic; corporate*
Text 2, Line 23-24	"Some wonder, therefore, whether directors of a corporation are permitted to optimize for stakeholders beyond the corporation's stockholders"	*Academic*
Text 2, Line 25	"This question is the subject of a rich debate, which we won't delve into here."	*Academic*
Text 1, Line 51-52	"and experimenting with education-centric programmes"	*Academic; scientific*
Text 2, Line 134	"we are empiricists"	*Scientific*
Text 2, Line 145-146	"on the theory that we'll"	*Academic*
Text 2, Line 4-5	"the graph goes up"	*Academic*

Understood from a legitimation perspective, the management literature on legitimacy within newly formed corporations provides multiple insights here. The AI industry is in its 'Introductory' phase which means it lacks legitimacy, thus requiring these "early entrants [to] devote a substantial amount of energy to sector building [through] creating objectivity and exteriority" (Suchman, 1995, p. 586). Understood through this lens, Anthropic, like OpenAI with their militaristic order of discourse, can be seen to pursue a "capturing" approach, meaning they use discourse as an associative activity to create new legitimating beliefs, a legitimation technique established by Reast et al. (2013, p. 148) in their modelling of legitimation strategies within controversial industries. Indeed, the utilisation of these orders of discourse mirrors the "subtle textual strategies" for corporate legitimation observed by Vaara and Tienari (2008, p. 988). From

an AI safety perspective, if corporate governance development is guided by principles of the scientific and academic methods, this could prove to be positive. However, viewed through this study's critical lens, purporting to be scientific and academic through discourse is not the same as prioritising these principles in practice.

4.4 Textual Analysis

In the textual analysis, almost every legitimation technique was observed throughout each of the four texts. The following analysis demarcates each broad legitimation technique for conceptual clarity, but it ought to be noted here that there is extensive overlapping between these categories. Mythopoeia, for instance, often involves moral evaluations and authoritative references; moral evaluation contains implicit goal-oriented rationalisation; authorisation requires some reference to institutions or individuals possessing morality; and rationalisations invariably make moral claims (van Leeuwen, 2007). As can be seen in the colour coded *Appendices*, varying legitimation techniques are employed often within the same sentence. Whilst the myriad themes and patterns that emerged in the analysis transcend these categories, the categories of Section 4 broadly ascertain to van Leeuwen's four categories, to enable greater accordance with the applied theory and to enhance comprehensibility (van Leeuwen & Wodak, 1999).

4.4.1 Ubiquitous Authorisation

Throughout the four texts, extensive evidence was found of actors using authorisation to legitimate themselves and their corporate structuring, by underlining their credibility regarding corporate performance, intelligence, AI and consideration of AI safety.

Most prominent were personal and expertise authorisation, two forms which ground legitimacy upon status in an institution or expertise on a certain topic, respectively (van Leeuwen, 2007, p. 94). This was done in multiple ways, including direct reference to the company's boards possessing expertise and insights (e.g. Text 2, Lines 92-93) but also through naming each board member along with their respective positions. Within Text 2 (Lines 153-157), Anthropic not only

name each board member but also name their statuses within their respective institutions. OpenAI (Text 1, Lines 108-110), also list their board members but "take for granted" the awareness of their board member's expertise, with them each being notable CEOs and heads of governmental institutions (van Leeuwen, 2007, p. 95). Situating this within the corporate governance literature, the emphasis on expertise perhaps ought to have been expected. Tylecotea & Ramirez (2006, p. 163), in their analysis of how different systems give rise to different corporate governance norms, found that 'outsider-dominated' systems, such as the U.S., "generate high industry-wide expertise [in their corporate governance structures]" and subsequent expectations of such.

Another dominant form of legitimation found in the analysis, that follows this same strategy, was the invocation of role model authorisation, in which an actor invokes an "opinion leader" (van Leeuwen, 2007, p. 95). The most notable example was found in Text 1, with OpenAI continually referencing "Microsoft", a corporation who have defined information technology. OpenAI takes this one step further by situating itself on an equal plane with Microsoft by using words and phrases that invoke reciprocity and mutual respect, like "partnership", "partner", and "why we chose Microsoft". From a legitimation theoretic perspective, in deploying personal, expertise and role model authorisation, these labs legitimate their corporate governance frameworks through borrowing legitimation from their respective board members and 'partnered' corporations, who maintain strong legitimation claims of their own. Aligning this with Reast et al.'s (2013, pp. 148, 144) research again suggests a 'capturing' strategy, with OpenAI effectively borrowing from Microsoft's legitimacy, as well as a 'construal' strategy, by "producing passive support and acquiescence".

Another form of authorisation observed was that of *informal authorisation*, through reference to laws and regulations. As can be seen in Table […], all four texts referred to legal constraints, laws, and the formulation of criteria by which they would subsequently abide, thus legitimating themselves by positioning their work as within the bounds of legal and normative regulation. However, there was a relatively small presence of this form of legitimation, in contrast to van Leeuwen & Wodak's (1999, p. 104) study on legitimation techniques within immigration policy, which found extensive reference to 'regulations' and 'the law'. This is likely due to the lack of regulatory frameworks currently active within AI safety (Cihon et al., 2020), a theme reflected in Text 2, Line 49-50, when Anthropic use predictive theoretical rationalisation to argue that AI will develop at a pace that renders laws and regulations insufficient for guaranteeing AI safety [See 4.4.4]. In undermining the efficiency of traditional formal regulation, they simultaneously attempt to legitimate the novel and untested nature of their LTBT corporate governance structures. This, in

turn, may represent a legitimation strategy described by Suchman (1995, p. 587) as an effort "to manipulate environmental structure by creating […] new legitimating beliefs."

Table 6: Varying Forms of Authorisation

Reference	***Quotes from Texts***	***Form of authorisation***
Text 1, Line 54	Collaboration with "Stanford University Artificial Intelligence Index Fund"	Personal; expertise; role-model
Text 1, Line 94	"our partnership with Microsoft includes a multibillion dollar investment"	*Personal; expertise; role model*
Text 1, Line 52	"OpenAI scholars"	*Expertise; personal*
Text 1, Lines 108-110	"Sam Altman, Adam D'Angelo, Dr. Sue Desmond-Hellmann, Retired U.S. Army General Paul M. Nakasone, Nicole Seligman, Fidji Simo, and Larry Summers."	*Expertise; personal; role-model;*
Text 2, Lines 153-157	"Jason Matheny: CEO of the RAND Corporation (Line 153-157) Kanika Bahl: CEO & President of Evidence Action Neil Buddy Shah: CEO of the Clinton Health Access Initiative (Chair) Paul Christiano: Founder of the Alignment Research Center Zach Robinson: Interim CEO of Effective Ventures US"	*Expertise; personal; role-model*
Text 2, Line 119-120	"the board will benefit from the insights of Trustees with deep expertise and experience"	*Expertise*
Text 2, Lines 169-172	"The Trust structure was designed and "red teamed" with immeasurable assistance by John Morley of Yale Law School, David Berger, Amy Simmerman, and other lawyers from Wilson Sonsini, and by Noah Feldman and Seth Berman from Harvard Law School and Ethical Compass Advisors."	*Personal; expertise*

Text 2, Lines 92-93	"Trustees with backgrounds and expertise in AI safety, national security, public policy, and social enterprise"	*Personal; expertise*
Text 3, Line 101	"different expertise that we want the board to have"	*Expertise*
Text 4, Line 26-27	"Now there are legal limits on that, of course"	*Impersonal*
Text 2, Line 80	"The legal latitude afforded by our PBC structure"	*Impersonal*
Text 3, Line 100	"we have some criteria that we think are important for the board to have"	*Impersonal*
Text 1, Line 40	"The for profit would be legally bound"	*Impersonal*
Text 2, Line 49-50	"The technology is advancing so rapidly that the laws and social norms that constrain other high-externality corporate activities have yet to catch up with AI;	*Rejection of impersonal authority*
Text 4, Line 4	""Even us who have not been super focused on commercialization and more on safety"	*Personal*

The analysis also yielded extensive evidence of a final form of authorisation named here as 'temporal' authorisation, which describes each actor making continual reference to time when legitimating their corporate governance structures. Whilst it could be construed as personal authorisation, van Leeuwen's (2007) work makes no explicit reference to time or temporal authorisation. Confronted by the untested nature of their corporate structures, they each make continued efforts to ground their structures as either having been developed for a long time or being built upon beliefs that the actors have maintained for a significant period. This is achieved using language like "beginning", "always", "harkening" and "birth", as well as marrying temporal authorisation with impersonal authorisation, found in Anthropic highlighting that their corporate governance structure has been "formally" developed since "the beginning".

Table 7: 'Temporal' Authorisation

Reference	***Quotes from Texts***	***Form of Authorisation***

Text 1, Line 9-11	"Since the beginning, we have believed that powerful AI [….] has the power to reshape society"	Personal
Text 1, Line 99	"From the beginning"	*Personal*
Text 1, Line 27	"We have always expected"	*Personal*
Text 1, Lines 52-53	"Over the years"	*Personal*
Text 1, Line 103	"Harkening back to our origins"	*Personal*
Text 4, Line 21-22	"some version of that has been in development since the beginning of Anthropic, even formally"	*Personal*
Text 2, Line 5	"we have been developing since the birth of Anthropic"	*Personal*
Text 3, Line 90	"I think I had even previous to that weekend suggested"	*Personal*

A further interesting example was provided by multiple actors representing both labs (e.g. Text 1, Line 9-11 and Text 4, Line 4) using temporal authority to underline their prioritisation of AI safety, thus suggesting that safety still remains at least a discursive priority. From an AI safety perspective, this could suggest that in at least some regard, the AI industry may be a "dual focus" industry, such as the airline industry, where profit and safety occupy a twinned priority (Gaba & Greve, 2019). If not, then this nonetheless represents a means by which legitimation is sought, as with oil companies seeking to legitimate through a CSR discourse (Du & Vieira, 2012).

In summarising the discussion of the wider implications of authorisation's ubiquitous presence, we ought to expect this technique's utilisation. The fundamental lack of inherited legitimacy within the AI industry, and the pace with which it is developing, results in actors seeking to legitimate as quickly as possible. "Capturing" and "construal" represent the optimal such means, with personal, impersonal and expertise legitimation efficient ways of doing so. The use of temporal authority represents a different strategy, with AI companies redefining their environment by manipulating public perceptions of how their experience. This finding supports the present literature on legitimation within emerging industries, as well as introduces a novel form of legitimation (temporal authorisation).

4.4.2 Company-Line Moralisation

The second dominant theme of the textual analysis was each company underlining its moral intentions, articulated centrally as a commitment to humanity, coupled with extensive moral abstraction and analogization of their corporate activities.

In addition to the moral evaluation highlighted through the various orders of discourse [Section 4.3.2], extensive instances of direct moral proclamations were found in Texts 1 and 2. Commitments to "the public good" and to creating AI that is "safe and benefits all of humanity", as well as words like "principles", "benefits, "safe", "guiding", "commitments", and "good", all represent moral evaluative adjectives, intended to reify their cultivated image of service and selflessness (van Leeuwen, 2007, p. 98). By aligning their goals with a moral valence, they frame their corporate activities as within the bounds of a strong moral framework, thus morally legitimating themselves and their corporate structures. This moral evaluation extended also to the companies' introduction of their board members, with Anthropic (Text 2, 160-161) listing the moral qualities that they prioritise in their board members.

Evident in this moral evaluation, in both Texts 1 and 2, is a strong stakeholder theoretic emphasis. The continued reference to responsibility extending beyond shareholders and employees is prototypical of stakeholder theory Ayuso et al. (2014), which is further evidenced by the fact that not once in a single of the four texts does an actor suggest allude to a sole fiscal responsibility to profit margins. This gives credence to Harris & Freeman's (2008) argument that the "separation thesis", which posits ethics can be separated from business, does not maintain a dominant hold within the AI corporate space. Through adopting this legitimation strategy, these companies discursively signal that they conceptualise themselves as having a greater responsibility beyond solely profit generation.

Patterns of morally evaluative legitimation can also be found in the naming and continued repetition of each company's respective corporate governance structures, which all bear strong moral connotations moral connotation. As alluded to in Section 4.2, OpenAI's published document focuses much of its attention on affirming its status as a Non-Profit, thus abstracting its corporate practices as altruistic rather than profit-seeking (van Leeuwen, 2007, p. 99). Anthropic's "Long-Term Benefit Trust" exhibits this also, through which they morally evaluate and abstract their corporate governance framework to instead be a 'Trust'. In doing so they further attempt to legitimate their corporate governance structures as prioritising AI safety, through framing their

structures as morally legitimate. A further form of moral abstraction was the repeated language found in both corporate documents of "pairing" and "partnerships", which both Anthropic and OpenAI use to abstract corporate procedures onto a moral plane, invoking ideas of a familial and moral relationship.

Table 8: Company-Line Moralisation

Reference	***Quotes from Texts***	***Form of Moral Evaluation***
Text 1, Line 18-19	"A project like this might previously have been the provenance of one or multiple governments""	*Analogy*
Text 1, Line 103	"this is a unique and ambitious project that requires resources at the scale of the public sector"	*Analogy*
Text 1, Line 50	"sponsoring a basic income study"	*Analogy*
Text 1, Line 23	"bound by strong commitments to the public good"	*Evaluation*
Text 1, Line 42	"guiding principles of safety and broad benefit"	*Evaluation*
Text 1, Line 6	"safe and benefits all of humanity"	*Evaluation*
Text 1, Line 52-53	"Over the years, the Nonprofit also supported a number of other public charities"	*Evaluation; abstraction*
Text 1, *20 mentions*	"OpenAI Nonprofit"	*Evaluation; abstraction*
Text 2, Line 11-12	"for the long-term benefit of humanity"	*Evaluation; abstraction*
Text 2, Line 74-76	"The public benefit purpose stated in Anthropic's certificate is the responsible development and maintenance of advanced AI for the long-term benefit of humanity."	*Evaluation; analogy*
Text 2, *19 mentions*	"Long-Term Benefit Trust" and "LTBT"	*Evaluation; analogy*
Text 2,*11 mentions*	"Public benefit" and "Public Benefit Corporation status"	*Evaluation; analogy*
Text 3, Line 62-63	"what we'd really like is for the board of OpenAI to answer to the world as a whole"	*Evaluation; abstraction*

Text 1, Lines 4-5	"a partnership between"	*Abstraction*
Text 2, Lines 10-11	"Paired with our"	*Abstraction*
Text 2, Lines 160-161	"to surface individuals who exhibit thoughtfulness, strong character, and a deep understanding of the risks, benefits, and trajectory of AI and its impacts on society"	*Evaluation*

Viewed from the perspective of Palazzo & Scherer's (2006), who advocate for a shift towards a morality-based, discourse-derived legitimacy, this extant moral evaluation of their corporate governance structures may represent a positive direction. However, from a critical legitimation perspective, it must remain clear that the moral evaluation is restricted at present to the realm of discourse. Parallels can be drawn here between CSR and AI safety legitimation, with there being evident tensions between legitimately socially responsible behaviour and companies simply cultivating their image (Moir, 2001). At present, neither corporate governance structure has endured sufficient stress testing for their moralising to be substantiated.

A final point to highlight here is the significant imbalance in moral evaluation between the companies and their respective CEOs. As seen in Table 8, as well as in the appendices, moral evaluation was predominantly found in the official companies' documents, rather than in the CEOs interviews, who instead legitimated themselves and their companies using mythopoeia [See Section 4.4.3]. This was perhaps in part owing to the companies having the ability to draft and redraft their documents [See Section 4.1], in contrast to the CEOs having to construct their narratives impromptu.

In bringing together the analysis and discussion of morally evaluative legitimation, we can first see a strong correlation between discursive CSR and AI safety legitimation. This is indicated by the extensive use of moral analogies and evaluations, as well as the word choice in the naming of these structures. Regarding the implications for AI safety here, the CSR literature would suggest cause for both optimism and concern. A discourse around safety and a duty to a greater responsibility is a two-sided coin: it is encouraging and an important part of any change, but, as with CSR, also represents a further opportunity for the manipulation of perceptions. Such concerns may be furthered by the imbalance between the labs' moral evaluation within their discourse and their CEOs, as the lesser of the latter perhaps enabling a truer perspective upon intentionality and safety priority.

4.4.3 CEOs' Hero-Driven Narratives and Cautionary Tales

Where the official corporate documents featured extensive moral evaluation by multiple means, the podcast interviews yielded a more dominant theme of mythopoeia, with each CEO drawing on narrative construction to legitimise themselves.

This theme was particularly apparent in Text 3, with Altman constructing a narrative of himself as the wronged hero, who subsequently has risen from the ashes. In discussing the November 2023 board saga, he first frames himself as a social actor harmed (Line 18-19, 26-27), thus villainising the board's ability to hire and fire CEOs, then asserts that "My company very nearly got destroyed" (Line 39-40). The language of "destruction" invokes adversarial violence, whilst his use of the possessive "My" further centres himself within the story. This narrative arc then culminates with the line "I felt like if I was going to come back […]" (Line 92), in which Altman emerges from the ashes, no longer viewing the board as an independent check on his power, but instead as subject to his demands. In doing so, he treads an interesting narrative line between cautionary and moral tales: he almost suffered wrongdoing, then nonetheless emerged victorious. To further legitimate these actions and himself, Altman draws on the militaristic discourse, in stating "You do that on the battlefield" (Line 98-99). In explaining Altman's legitimation strategies using the CG literatures' contributions, Altman's personal insertion into the narrative could be attributed to the greater personal involvement of founder-CEOs, who are psychologically more emotionally involved with their company (Wasserman, 2006).

A further interesting phenomenon is observed in the inconsistent internal logic of Altman's narratives, shown by his concurrent delegitimation of the board's expertise and his use of them to legitimate the new board. He first undermines their personal and expertise authority, both explicitly (Text 3, Line 78-79, 50) and implicitly (Text 3, Line 50), thus casting them as an antagonist. Moments later however, (Text 3, Line 88-89), Altman uses the judgement and expertise of the previous board members to legitimise the new board members. Whilst the board members' 'outsider' status may have meant they lacked understanding of the company and requisite knowledge about R&D efficacy (Bertoni et al., 2013, p. 373) this doesn't explain the inconsistencies in his narrative construction. However, precedent for this is found in the management literature, with Tienari et al.'s (2003) analysis of how corporate actors utilise narrative finding that individual actors often draw on conflicting narratives at different points in time.

Table 9: CEO Mythopoeia

Reference	***Quotes from Texts***	***Form; Narrative***
Text 3, Line 18-19	"really unpleasant and really difficult"	*Cautionary/moral; Altman persona*
Text 3, Line 26-27	"really painful and hard"	*Cautionary/moral; Altman personal*
Text 3, Line 39-40	"My company very nearly got destroyed"	*Cautionary/moral; Altman personal*
Text 3, Line 92	"I felt like if I was going to come back, I needed new board members"	*Cautionary/moral; Altman personal*
Text 3, Line 98-99	"You do that on the battlefield. You don't have time to design a rigorous process then"	*Cautionary/moral; Altman personal*
Text 3, Line 78-79	"the board also I think didn't have a lot of experienced board members"	*Cautionary/moral; Board delegitimation*
Text 3, Line 52	Infers the board made "suboptimal decisions"	*Cautionary/moral; Board delegitimation*
Text 3, Line 50	"the board members are well-meaning people on the whole"	*Cautionary/moral; Board delegitimation*
Text 3, Line 88-89	""And we were trying to agree on new board members that both the executive team here and the old board members felt would be reasonable"."	*Cautionary/moral; Narrative inconsistencies*
Text 4, Line 64-66	"attaching your incentives very strongly to the approval or cheering of a crowd can destroy your mind, and in some cases, it can destroy your soul"	*Cautionary; Amodei's learnings*
Text 4, Line 69-70	"I've seen cases of folks who are deep learning sceptics, and they become known as deep learning sceptics on Twitter"	*Cautionary; Amodei's learnings*

In Text 4, Amodei also uses mythopoeia but to a different degree. Unlike Altman, he does not explicitly insert himself into the narrative, instead constructing a sage-like cautionary tale. This

is most evident in his precaution, “attaching your incentives very strongly to the approval or cheering of a crowd can destroy your mind, and in some cases, it can destroy your soul” (Line 64-66). By not using first person pronouns Anthropic’s CEO here maintains his academic order of discourse, but implies that he has avoided this ill-fate. This cautionary tale draws upon the age-old narrative archetype of the character who pursues lofty ambitions, and often succeeds, but loses themselves during the pursuit.

Returning to the contextual analysis, the differences in the CEO’s approaches could be attributed to the November 2023 board saga. In undermining both the legitimacy of Altman and OpenAI, as well as their corporate governance structure, the labs and their respective leaders were given significantly different positions to legitimate from. Where Amodei takes a more abstract, sage-like tone within his cautionary tales, Altman instead inserts himself as a central character. Suchman (1995, p. 598), in discussing legitimation strategies, highlights that CEOs may attempt “to justify the disruption […] to make the disruptive events appear consonant with prevailing moral and cognitive beliefs” when legitimating from a damaged perspective, in contrast to the “far easier enterprise” of legitimating when perceived well. As such, this may explain Altman’s ego-centric mythopoeic endeavours, in contrast to Amodei’s selfless mythopoeic constructions.

In grouping and clarifying the discussion points raised here, we can see that the clear contrast between the mythopoeic legitimation strategies of each respective CEO can likely be attributed to their contextual histories. This finding contributes further to Beelitz and Merkl-Davies’ (2011, p. 115) work on CEO discourse manipulation, which found narrative construction to be a valuable tool in manufacturing “organisational audiences’ consent in the wake of crises, with Altman’s crisis furthering the use of this technique. These findings also demonstrate the deep personal attachment that Altman feels towards OpenAI coupled with Altman’s framing himself as a wronged victim and an adversary of the board, which may signal negative news for AI safety. Conceiving of himself in this way may cloud his judgement in prioritising what is best for wider stakeholders, whilst also negatively warping his perspective on building corporate governance structures that prioritise safety, even if that would mean him being removed from the company.

4.4.4 Rationalization Abound

Regarding rationalisation strategies, two dominant themes emerged in each actor’s respective legitimation attempts, both of which served as rationalisations of their actions, themselves, and their governance structures.

The first is described here as the extensive use of theoretical rationalisations by all four actors, in definitory, explanatory, and predictive forms. Through framing corporate governance norms and their own actions in specific ways, they differentially construct social reality relating to AI, corporate governance, and their own legitimacy (van Leeuwen, 2007, p. 101). For instance, by defining AGI (Text 1, Line 67-69), OpenAI makes a series of legitimating moves. First, they shape people's perceptions of what AGI is, second, they fail to contextualise their definition, instead framing it as an objective truth, and third, the definition is sufficiently vague to allow significant revisions. In using definitions in this manner, they simultaneously legitimate their actions, by using theoretical rationalisation to define reality in a way that benefits them, whilst also giving themselves sufficient definitional manoeuvrability that in turn serves to undermine their structures. As OpenAI's safety strategy includes the exclusion of AGI from commercial deals, this discursive manipulation of definitional parameters could potentially negatively impact the safeguarding of AI. The weaving together of these different forms of rationalisation is also observed in Anthropic's legitimation strategies, as can be seen in the further examples in Table 10.

Table 10: Theoretical Rationalisations

Reference	***Quotes from Texts***	***Form of Rationalisation***
Text 1, Line 81	"when we've attained AGI"	TR prediction
Text 1, Line 35-55	"The OpenAI Nonprofit would remain intact […]"	TR explanation
Text 1, 67-69	"We enacted this by having the Nonprofit wholly own and control a manager entity (OpenAI GP LLC) that has the power to control and govern the for-profit subsidiary."	TR definition
Text 1, Line 67-69 (& Line 81-82)	"culminating in AGI—meaning a highly autonomous system that outperforms humans at most economically valuable work"	TR definition
Text 2, Line 31-33	"At Anthropic, our perspective is that the capacity of corporate governance to produce socially beneficial outcomes depends strongly on non-market externalities.	TR explanation
Text 2, Line 33-35	"Externalities are a type of market failure that occurs when a transaction between two parties imposes costs or benefits on a third party who has not consented to the transaction."	TR explanation & definition

Text 2, Line 46-47	"We believe AI may create unprecedented externalities"	TR prediction
Text 2, Line 58	"We do not expect"	TR prediction
Text 2, L14-15	"A corporation is overseen by its board of directors. The board selects and oversees the leadership team (especially the CEO), who in turn hire and manage the employees."	TR definition
Text 2, Line 53-54	"To be clear, for most of the day-to-day decisions Anthropic makes, public benefit is not at odds with commercial success or stockholder returns"	TR definition; IR goal-oriented
Text 3, Line 57-58	"one of the things that we did see is in most corporate structures, boards are usually answerable to shareholders."	TR explanation
Text 2, Line 41-43	"But other parties, such as the general public, don't directly contract with a corporation and therefore do not have a means to charge or pay for the costs and benefits they experience."	TR explanation
Text 1, L35-55	*OpenAI's account of why and how the non-profit retains control over the for-profit arm*	TR (all forms) and IR (both forms)
Text 3, L120-122	"And then you need some people who are like, "How can we deploy this in a way that will help people in the world the most?" And	IR goal-oriented
Text 4, L24-26	"Every traditional investor who invests in Anthropic looks at this […]Some of them are like, oh my god, this body of random people could move Anthropic in a direction that's totally contrary to shareholder value"	TR explanation
Text 4, L76-77	"Everyone wants a friendly face, but actually, friendly faces can be misleading."	TR explanation, prediction

The second core theme observed, as related to rationalisation, was the extensive reference to goal-oriented instrumental rationalisation. As observed in Table 11, all actors legitimated their corporate governance frameworks in terms of goal-oriented rationalisations. Consistent reference is made by both Anthropic and OpenAI, to how their governance frameworks are constructed (thus referencing the means) and how they serve in meeting their goals outlined (thus, the ends), which

are generally related to benefitting humanity. One interesting example was Altman's goal-oriented rationalisation (Text 1, Lines 104-105 and 120-122), in which he outlined his priorities for board expertise, as well as for what "the board needs to [do]" (Text 3, Line 102). At first, he centred on the goal of having a board that maintains expertise related solely to traditional corporate governance duties and only highlighted AI safety credentials as a priority when prompted to do so by the interviewer. Anthropic and Amodei, contrarily, focus instead on continually highlighting that understanding of AI safety is an essential credential for the LTBT's trustees.

Discussing the potential implications, it could be inferred that Altman does not believe safety credentials are as necessary for a board member as corporate expertise, which could well owe to the differing corporate structures Anthropic and OpenAI maintain [Section 4.2]. In having one board, OpenAI are faced with the difficult task of balancing corporate and safety priorities within one body, whereas the LTBT enables Anthropic to have one board focused on safety, whilst the other can focus on corporate concerns. Regarding the wider implications for AI safety, OpenAI's lack of divisions of power could create further problems down the line. A further point to raise here is the essentiality of maintaining a safety rhetoric [seen in Section 4.4.1], as demonstrated by Altman revising his original prioritisation of board members' qualifications.

Table 11: Instrumental Rationalisation

Reference	***Quotes from Texts***	***Form of Rationalisation***
Text 1, L33-34	"So we devised a structure to preserve our Nonprofit's core mission, governance, and oversight while enabling us to raise the capital for our mission"	*IR goal-oriented*
Text 3, L104-105	"thinking about a group of people that will bring nonprofit expertise, expertise at running companies, good legal and governance expertise, that's what we've tried to optimize for."	*IR goal-oriented*
Text 3, Line 120-122	"Look, I think you definitely need some technical experts there. And then you need some people who are like, "How can we deploy this in a way that will help people in the world the most?"	*IR goal-oriented*

Text 4, Line 67-68	"I've deliberately tried to be a little bit low profile because I want to defend my ability to think about things intellectually"	*IR goal-oriented*
Text 3, Lines 41-42	"how to build a resilient org and how to build a structure that will stand up to a lot of pressure in the world, which I expect more and more as we get closer"	*IR goal-oriented; predictive TR*
Text 1, Lines 42-43	"Throughout, OpenAI's guiding principles of safety and broad benefit would be central to its approach.	*IRl, means-oriented*
Text 1, L 73	"The Nonprofit's principal beneficiary is humanity, not OpenAI investors."	*IR goal-oriented (stakeholder emphasis)*
Text 1, L 45-47	"to incentivize them to research, develop, and deploy AGI in a way that balances commerciality with safety and sustainability, rather than focusing on pure profit-maximization.	*IR means-oriented (agentic theoretic) for IR goal-oriented (stakeholder)*
Text 2, L109-111	"The Trust must use its powers to ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic's conduct and our public benefit purpose."	*IR Goal-oriented (stakeholder theoretic)*
Text 4, L74-76	"I think it distracts people […] I want people to think in terms of the nameless, bureaucratic institution and its incentives"	*IR means-oriented (agentic theoretic emphasis)*
Text 2, L85-86	"We set out to design a structure that would supply our directors with the requisite accountability and incentives	*IR goal-oriented (agentic theoretic emphasis)*
Text 2, L118-119	"and will thus have incentives to appropriately balance the public benefit with stockholder interests."	*IR means-oriented (agentic theoretic) for goal-oriented (stakeholder theoretic)*
Text 2, Line 41-43	"But other parties, such as the general public, don't directly contract with a corporation and therefore do not have a means to charge or pay for the costs and benefits they experience."	*Explanatory TR, Agency theoretic emphasis,*

A final interesting trend was the consistent allusion to both stakeholder and agency-theoretic corporate governance frameworks, with the former typically constituting the end, and the latter constituting the means. OpenAI underlines that the non-profit is created to benefit the public (i.e. wider stakeholders) rather than investors (Text 1, Line 73), whereas Anthropic continually reference the importance of balancing profits with a duty to greater public benefit (Text 2, L118-119). Conversely, both companies' framing of the workings of corporate governance mechanisms is often framed through the lens of agency theory. As can be seen in Table 11, continual reference is made to "incentives", which is a core tenet of agency theory, as well as to "contracts" (Jensen & Meckling, 2000). For instance, as seen in Text 1, Line 45-47, OpenAI uses goal-oriented rationalisation to articulate their stakeholder-theoretic goal (to "research, develop, and deploy AGI in a way that balances commerciality with safety and sustainability, rather than focusing on pure profit-maximization) through agency-theoretic means (by focusing on incentives") investors and AI-safety oriented observers.

To summarise the key observations surrounding rationalisations, both theoretic and instrumental rationalisations were used extensively by all actors to discursively construe their own conceptions of AI and to justify their actions to both stakeholders and investors. Regarding wider discussion points, stakeholder and agentic theoretic perspectives were invoked, providing further insight into how these companies conceptualise themselves, whilst Altman's rationalisations created further distance between himself and Amodei, and signalled potential pitfalls of OpenAI's corporate structuring.

Chapter 5: Conclusion

As has been demonstrated through the undertaking and interweaving of the dialectical [4.1], contextual [4.2], intertextual [4.3] and textual analyses [4.4], there exists innumerable rich, complex and consequential links between the concepts of AI safety, legitimation and corporate governance. So novel is the nature of their corporate governance structures, and so high are the stakes surrounding the development of AGI, that AI corporations' efforts to legitimate require abundant legitimation techniques. This study has shown that "meaning making" nature of discourse offers a powerful means of legitimation which AI companies are making significant use of (Fairclough, 2010, p. 3).

5.1 Answering the Research Questions

Turning back to the research questions posed at the commencement of this study, which were then open-ended and undetermined, we can now provide answers to them which provide us with an increased perspective.

> *RQ1: What are the textual and intertextual legitimation techniques that leading AI labs use to legitimate their novel corporate governance structures?*

Throughout their discursive legitimation, both the corporations themselves and their respective CEOs utilised almost the entirety of van Leeuwen's (2007) repertoire of legitimation techniques. Authorisation in all its forms was universally utilised, including the use of 'temporal authorisation', to legitimate their respective beliefs, attributes, status and conformance with regulation. Moral evaluation was found to represent a primary means of legitimation within these companies' official documents, being expressed through textual choices as well as drawing on varied orders of discourse [4.2]. Mythopoeia was entirely restricted to the use of the CEOs, although with varying emphases, and rationalisation, both theoretical and instrumental, was used by all actors to construe their goals and their conceptions of social reality as legitimate.

RQ2: How do leading AI labs conceptualise their responsibilities in relation to themselves and corporate governance?

The extensive legitimation techniques deployed point to the conclusion that these corporations are undoubtedly strategic actors, but also that they do conceptualise of themselves as having greater responsibilities to wider stakeholders. This was evidenced through the orders of discourse utilised [Section 4.2], the extensive moral evaluation [Section 4.4.2], and the evident stakeholder theoretic emphasis found in their rationalisations [Section 4.4.4]. Using discursive strategies, they at least present themselves to be benevolent guardians, although ego-centric perspectives [Section 4.4.3] and potentially insecure CG structures [Section 4.2] may give reason to think otherwise.

RQ3: What wider implications do these findings have for AI safety?

Regarding wider implications for AI safety, it is evident that concern for this subject is both evident in these discourses, but also that it is being used as a legitimative technique [Section 4.4.1]. The corporate governance structures analysed here [Section 4.2] represent varying levels of comprehensiveness, with the many competing interests and involved actors further complicating the prioritisation of AI safety, whilst the implications of the militaristic order of discourse provide further concern. Multiple legitimation strategies appeared to lack honesty and weight, thus furthering concerns that the structures currently in place are not sufficient for safeguarding humanity's future.

5.2 Contributions, Recommendations, and Future Research Directions

This study has made significant contributions to knowledge and theory, in both meaningful and minor ways. Most centrally, it represents the first effort to bring together the three literatures of corporate governance, AI safety and legitimation, providing novel insights not previously found in any of these literatures in the process. Utilising a CDA that combined both van Leeuwen's (2007) and Fairclough's (2010) frameworks, provides a minor methodological contribution by demonstrating the utility of their synthesis. Finally, in highlighting 'temporal authorisation' as a

novel differentiated legitimation technique, this study gives greater depth to van Leeuwen's legitimation theoretic framework, thus representing a minor theoretical contribution.

Regarding recommendations for practice, it is recommended that policymakers explore the possibility of stricter testing, appraisal and regulation of AI corporations' corporate governance structures, with discursive legitimation alone not representing a sufficiently safe means by which to judge their aptitude. For organisations seeking to develop AI, the recommendation is simple: when discursively legitimating, practice maximal self-awareness regarding the organisations' CG weaknesses, to enable honest communication and the open critique of dangerous practices.

The previously lacking contribution from the management literature to this issue [2.1.3] combined with the prescience of this topic, results in extensive possibilities for future research. From a legitimation standpoint, analyses of different AI labs would provide further illumination, exploring the differences between private and publicly traded AI corporations. From a corporate governance perspective, there exists multiple pressing areas of research, including both quantitative and qualitative analyses of the varying board structures, as well as analyses of the backgrounds and representations of individual board members, for instance, comparing OpenAI's board before and after November 2023. Furthermore, research on these topics from agentic and stakeholder theoretic perspectives would bring further theoretic light and unification of the AI safety and corporate governance literature.

5.3 Final Remarks

Returning to the quote with which this study opened, Altman's prophesising may well be validated in his predictions of a power struggle. Who this power resides in the hands of and what they choose to do with it could have profound impacts on all areas of our lives. As such, the need for precise and critical analyses of these strategies is essential for illuminating where this power lies and how it is utilised by varying actors to manipulate public perceptions and legitimate themselves and their actions. This paper has sought to lay some modest foundations in advocating for and contributing to these issues, with the hope of inspiring further research. Contributions from management scholars to these topics in the coming years and decades are essential for improving not only our academic knowledge base but also our chances of ensuring this technology really is developed for the benefit of all humankind.

Bibliography

Abdeldayem, M., Aldulaimi, S., Abu-AlSondos, I., & Baqi, A. (2023). Corporate Governance and Sustainability Development Goals: Boeing Case Study. *Conference on Sustainability and Cutting-Edge Business Technologies* (pp. 354-366). Cham: Spinger Nature Switzerland.

Aldrich, H. E., & Fiol, C. M. (1994). Fools Rush in? The Institutional Context of Industry Creation. *Academy of Management Review, 19*(4), 645-670.

Andhov, A. (2024). OpenAI's Transformation: From a Non-profit to a 100 Billion Valuation. *Available at SSRN 4750197*.

Anthropic. (2023, September 19). *The Long-Term Benefit Trust*. Retrieved from Anthropic: https://www.anthropic.com/news/the-long-term-benefit-trust

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI & Society, 31*, 201-206.

Askell, A., Brundage, M., & Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. *arXiv*. doi:1907.04534

Ayuso, S., Rodríguez, M., García-Castro, R., & Ariño, M. (2014). Maximizing Stakeholders' Interests: An Empirical Analysis of the Stakeholder Approach to Corporate Governance. *Business & Society, 53*(3), 414-439.

Bainbridge, S. M. (2008). *The New Corporate Governance in Theory and Practice.* Oxford: Oxford University Press.

Band, D. (1992). Corporate governance: Why agency theory is not enough. *European Journal of Management, 10*(4), 453-459.

Bansal, T. (2024, January 16). *Which Company Will Ensure AI Safety? OpenAI Or Anthropic*. Retrieved from Forbes: https://www.forbes.com/sites/timabansal/2024/01/16/openai-or-anthropic-which-will-keep-you-more-safe/

Barnett, M. L. (2019). The Business Case for Corporate Social Responsibility: A Critique and an Indirect Path Forward. *Business & Society, 58*(1), 167-190.

Barrat, J. (2023). *Our final invention: Artificial intelligence and the end of the human era.* UK: Hachette.

Baysinger, B., & Butler, H. (2019). Corporate governance and the board of directors: Performance effects of changes in board composition. *Corporate Governance*, 215-238.

Bebchuk, L., Cohen, A., & Ferrell, A. (2009). What matters in corporate governance? *The Review of Financial Studies, 22*(2), 783-827.

Beelitz, A., & Merkl-Davies, D. M. (2011). Using Discourse to Restore Organisational Legitimacy: 'CEO-speak' After an Incident in a German Nuclear Power Plant. *Journal of Business Ethics*, 101-120.

Berger, P. L., & Luckmann, T. (1967). *The Social Construction of Reality.* New York: Open Road.

Bertoni, F., Colombo, M. G., & Croce, A. (2013). Corporate Governance in High-Tech Firms. In M. Wright, & e. al., *The Oxford Handbook of Corporate Governance* (pp. 365-38). Oxford: Oxford University Press.

Bhagat, S., & Bolton, B. (2008). Corporate Governance and Firm Performance. *Journal of Corporate Finance*, 257-273.

Biloslavo, R., Bagnoli, C., Massaro, M., & Cosentino, A. (2020). Business model transformation toward sustainability: the impact of legitimation. *Management Decision, 58*(8), 1643-1662.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines, 22*, 71-85.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford: Oxford University Press.

Breeze, R. (2012). Legitimation in corporate discourse: Oil corporations after Deepwater Horizon. *Discourse and Society, 23*(1), 3-18.

Bruno, V., & Claessens, S. (2010). Corporate governance and regulation: Can there be too much of a good thing? *Journal of Financial Intermediation, 19*(4), 461-482.

Cheffins, B. (2013). The History of Corporate Governance. In M. Wright, & e. al., *The Oxford Handbook of Corporate Governance* (pp. 46-64). Oxford: Oxford University Press.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of Humanity Institute. University of Oxford*, 340-342.

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society, 2*(4), 200-209. doi:10.1109

Cihon, P., Maas, M. M., & Kemp, L. (2020). Should artificial intelligence governance be centralised? Design lessons from history. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (pp. 228-234).

Cihon, P., Schuett, J., & Baum, S. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information, 12*(7), 275.

Clark, J. A. (2019). Regulatory markets for AI safety. *arXiv preprint arXiv*. doi:2001.00078

Cox, J., Martinez, E., & Quinlan, K. (2008). Blogs and the corporation: managing the risk, reaping the benefits. *Journal of Business Strategy, 29*(3), 4-12.

Critch, A., & Krueger, D. (2020). AI Research Considerations for Human Existential Safety (ARCHES). *arXiv*. doi:2006.04948

Cummings, B. (2012). Benefit corporations: How to enforce a mandate to promote the public interest. *Columbia Law Review, 112*, 578.

Deephouse, D., Bundy, J., Tost, L., & Suchman, M. (2017). Organizational Legitimacy: Six Key Questions. In R. Greenwood, C. Oliver, T. B. Lawrence, & R. E. Meyer, *The SAGE Handbook of Organizational Institutionalism* (pp. 27-52). Sage Publications Ltd.

Demers, C., Giroux, N., & Chreim, S. (2003). Merger and acquisition announcements as corporate wedding narratives. *Journal of Organizational Change Management, 16*(2), 223-242.

Donaldson, T., & Preston, L. (1995). The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications. *The Academy of Management Review, 20*(1), 65-91.

Du, S., & Vieira, E. (2012). Striving for Legitimacy Through Corporate Social Responsibility: Insights from Oil Companies. *Striving for legitimacy through corporate social responsibility: Insights from oil companies, 110*, 413-427.

Dubber, M. D., Pasquale, F., & Das, S. (2020). *Oxford Handbook of Ethics of AI.* Oxford: Oxford University Press.

Eleftheriadis, I., & Anagnostopoulou, E. (2015). Relationship between Corporate Climate Change Disclosures and Firm Factors. *Business Strategy and the Environment, 24*(8), 780-789.

EU Artificial Intelligence Act. (2024, April 2). *Home Page*. Retrieved from EU Artificial Intelligence Act: https://artificialintelligenceact.eu/

Fairclough, N. (1992). Discourse and text: linguistic and intertextual analysis within discourse analysis. *Discourse and Society, 3*(2), 193-217.

Fairclough, N. (2010). *Critical Discourse Analysis: The Critical Study of Language* (Second ed.). London: Routledge.

Fairclough, N. (2017). CDA as dialectical reasoning. In J. Flowerdew, & J. Richardson, *Flowerdew, J.; Richardson, J.E.* (pp. 35-51). London: Routledge.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., . . . Maple, C. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence volume, 3*, 566-571.

Freeman, R., Harrison, J., Wicks, A., Parmar, B., & De Colle, S. (2010). *Stakeholder Theory: The state of the art.* Cambridge: Cambridge University Press.

Fridman. (2007). The Social Responsibility of Business Is to Increase Its Profits. In Heidelberg, *Corporate ethics and corporate governance* (pp. 173-178). Berlin: Springer.

Fridman, L. (2024, 24 March) *Transcript for Sam Altman: OpenAI, GPT-5, Sora, Board Saga, Elon Musk, Ilya, Power & AGI | Lex Fridman Podcast #419*. Retrieved from: https://lexfridman.com/sam-altman-2-transcript/

Friedman, A., & Miles, S. (2002). Developing stakeholder theory. *Journal of management studies, 39*(1), 1-21.

Gaba, V., & Greve, H. (2019). Safe or Profitable? The Pursuit of Conflicting Goals. *Organization Science, 30*(4), 647-667.

Glozer, S., Caruana, R., & Hibbert, S. (2019). The never-ending story: Discursive legitimation in social media dialogue. *Organization Studies, 40*(5), 625-650.

Golato, A. (2017). Naturally Occurring Data. In A. Barrron, Y. Gu, & G. Steen, *The Routledge handbook of pragmatics* (pp. 21-26). Routledge.

Gompers, P., Ishii, J., & Metrick, A. (2003). Corporate governance and equity prices. *The quarterly journal of economics, 118*(1), 107-156.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research, 62*, 729-754.

Grix, J. (2002). Introducing Students to the GenericTerminology of Social Research. *Politics, 22*(3), 175-186.

HAI. (2024). *The AI Index Report: Measuring Trends in AI.* Stanford: Stanford University.

Harris, J., & Freeman, R. (2008). The Impossibility of the Separation Thesis: A Response to Joakim Sandberg. *Business Ethics Quarterly, 18*(4), 541-548.

Hart, S. (2010). Self-regulation, Corporate Social Responsibility, and the Business Case: Do they Work in Achieving Workplace Equality and Safety? *Business Ethics, 90*, 585-600.

Heizmann, H., & Fox, S. (2019). O Partner, Where Art Thou? A critical discursive analysis of HR managers' struggle for legitimacy. *International Journal of Human Resource Management, 30*(13), 2026-2048.

Hiller, J. (2013). The benefit corporation and corporate social responsibility. *Journal of Business Ethics*, 287-301.

IMGBIN. (2019, January 13). *Ouroboros*. Retrieved from IMGBIN: https://imgbin.com/png/4kiE59Dx/snake-the-cosmic-serpent-ouroboros-tail-eating-png

Jäger, S. (2011). Discourse and Knowledge: Theoretical and Methodological Aspects of A Critical Discourse and Dispositive Analysis. In R. Wodak, & M. Meyer, *Methods of Critical Discourse Analysis* (pp. 32-59). Newbury Park: SAGE Publications.

Jarrett, K. (2009). Private Talk in the Public Sphere: Podcasting as Broadcast Talk. *Communication, Politics & Culture, 42*(2), 116-135.

Jenkins, H. (2004). A Critique of Conventional CSR Theory: An SME Perspective. *Journal of General Management, 29*(4), 37-57.

Jensen, M. C., & Meckling, W. H. (2000). Theory of the Firm: Managerial Behaviour, Agency Costs and Ownership Structures. In K. RS, & P. L, *The Economic Nature of the Firm: A Reader* (pp. 283-303). Cambridge: Cambridge University Press.

Johnston, J. W. (2023). A Case for AI Safety via Law. *arXiv*. doi:2309.12321

Jones, T. (1980). Corporate social responsibility revisited, redefined. *California management review, 22*(3), 59-67.

Juric, M., Sandic, A., & Brcic, M. (2020). AI safety: state of the field through quantitative lens. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1254-1259.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., . . . Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*. doi:2001.08361

Kim, B., Hong, S., & Cameron, G. (2014). What Corporations Say Matters More than What They Say They Do? A Test of a Truth Claim and Transparency in Press Releases on Corporate Websites and Facebook Pages. *Journalism and Mass Communication Quarterly, 91*(4), 811-829.

Knauth, D. (2024, February 22). *Crypto exchange FTX to sell shares in AI startup Anthropic*. Retrieved from Reuters: https://www.reuters.com/technology/crypto-exchange-ftx-sell-shares-ai-startup-anthropic-2024-02-22/

Kurland, N. (2017). Accountability and the public benefit corporation. *Business Horizons, 60*(4), 519-528.

LaGrandeur, K. (2021). How safe is our reliance on AI, and should we regulate it? *AI Ethics, 1*, 93-99.

Latin, H. (1988). Good science, bad regulation, and toxic risk assessment. *Yale Journal on Regulation, 5*, 89.

Lazonick, W., & O'Sullivan, M. (1996). Organization, Finance and International Competition. *Industrial and Corporate Change, 5*(1), 1-49.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI in the public sector.* The Alan Turing Institute. doi:10.5281/zenodo.3240529

Lieberman, M. B., & Montgomery, D. B. (1988). First-mover advantages. *Strategic management journal, 9*, 41-58.

Love, J., & Hubbard, T. (2009). Prizes for Innovation of New Medicines and Vaccines. *Annals of Health Law, 18*, 155-186.

Lubatkin, M. (2007). One More Time: What Is a Realistic Theory of Corporate Governance? *Journal of Organisational Behaviour, 28*(1), 59-67.

Lund, D. S., & Pollman, E. (2021). The Corporate Governance Machine. *Columbia Law Review*, 2563.

Luyckx, J., & Janssens, M. (2016). Discursive legitimation of a contested actor over time: The multinational corporation as a historical case (1964-2012). *Organization Studies, 37*(11), 1595-1619.

Maas, M. (2022). Aligning AI regulation to sociotechnical change. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson, *Oxford Handbook on AI Governance (forthcoming)* (p. N/A). Oxford: Oxford University Press.

Magma, M. T. (2023). *A New Era of Influence: Podcasters' Emergence as One of Today's Most Influential Figures in Media.* Vox Media.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46-60.

Marchant, G., & Wallach, W. (2013). Governing the governance of emerging technologies. In G. E. Marchant, & K. W. Abbot, *Innovative Governance Models for Emerging Technologies* (pp. 136-155). Cheltenham: Edward Elgar Publishing Company.

McKinsey. (2023). *The State of AI in 2023: Generative AI's Breakout Year.* New York: Quantum Black AI by McKinsey.

Meyer, M. (2011). Between Theory, Method, and Politics: Positioning of the Approaches to CDA. In R. Wodak, & M. Meyer, *Methods of Critical Discourse Analysis* (pp. 14-30). Newbury Park: SAGE Publications.

Miles, S. (2017). Stakeholder theory classification: A theoretical and empirical evaluation of definitions. *Journal of Business Ethics, 142*, 437-459.

Moir, L. (2001). What do we mean by corporate social responsibility? *Corporate Governance, 1*(2), 16-22.

Monks, R. A., & Minow, N. (2011). *Corporate Governance.* John Wiley and Sons.

Morsing, M., & Schultz, M. (2006). Corporate social responsibility communication: stakeholder information, response and involvement strategies. *Business Ethics: A European Review, 15*(4), 323-338.

Mouton, C. A., Lucas, C., & Guest, E. (2024). *The Operational Risks of AI in Large-Scale Biological Attacks.* Santa Monica: RAND Corporation.

Naciti, V., Cesaroni, F., & Pulejo, L. (2022). Corporate governance and sustainability: a review. *Journal of Management and Governance, 26*, 55-74.

Nichols, T., & Walters, D. (2017). *Safety or Profit? International Studies in Governance, Change and the Work Environment.* Routledge: Oxon.

OpenAI. (2024, June 5). *Our Structure*. Retrieved from OpenAI: https://openai.com/our-structure/

Osofsky, H. (2011). Multidimensional governance and the BP deepwater horizon oil spill. *Florida Law Review, 63*, 1077.

Palazzo, G., & Scherer, A. (2006). Corporate Legitimacy as Deliberation: A Communicative Framework. *Journal of Business Ethics, 66*, 71-88.

Patel, D. (2023, August 08). *Dario Amodei (Anthropic CEO) - Scaling, Alignment, & AI Progress*. Retrieved from https://www.dwarkeshpatel.com/p/dario-amodei

Phillips, N., & Hardy, C. (2002). *Discourse Analysis: Investigating Processes of Social Construction* . London: Sage Publications.

Phillips, R., Freeman, R., & Wicks, A. (2003). What Stakeholder Theory is Not. *Business Ethics Quarterly, 13*(4), 479-502.

Quintelier, K., & Vock, M. (2022). The effect of mixing stakeholder value and profit on cooperation: You can't have your cake and eat it too. *European Management Journal*.

Raelin, J., & Bondy, K. (2013). Putting the good back in good corporate governance: The presence and problems of double-layered agency theory. *Corporate Governance: An International Review*, 420-435.

Raheja, C. (2005). Determinants of board size and composition: A theory of corporate boards. *Journal of financial and quantitative analysis, 40*(2), 283-306.

Rashid, A. (2021). Board independence and corporate social responsibility reporting: mediating role of stakeholder power. *Management Research Review, 44*(8), 1217-1240.

Reast, J., Maon, F., Lindgreen, A., & Vanhamme, J. (2013). Legitimacy-Seeking Organizational Strategies in Controversial Industries: A Case Study Analysis and a Bidimensional Model. *Journal of Business Ethics, 118*, 139-153.

Roose, K. (2023, February 3). *How ChatGPT Kicked Off an A.I. Arms Race*. Retrieved from The New York Times: https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html

Russell, S. J. (2022). *Artificial Intelligence: A Novel Approach* (4th ed.). Harlow: Pearson Education Ltd.

Sarkar, S., & Searcy, C. (2016). Zeitgeist or chameleon? A quantitative analysis of CSR definitions. *Journal of Cleaner Production, 135*(1), 1423-1435.

Saunders, M., Lewis, P., & Thornhill, A. (2023). *Research Methods for Business Students* (9th ed.). Harlow: Pearson.

Segal, L. (2017). Benefit Corporations: A Step towards Reversing Capitalism's Crisis of Legitimacy. *Va. J. Soc. Pol'y & L., 24*, 97.

Singh, J., Tucker, D., & House, R. (1986). Organizational Legitimacy and the Liability of Newness. *Administrative Science Quarterly, 31*(2), 171-193.

Sotala, K. (2018). Disjunctive scenarios of catastrophic AI risk. *Artificial intelligence safety and security*, 315-337.

Suchman, M. C. (1995). Managing Legitimacy: Strategies and Institutional Approaches. *Academy of Management Review, 20*(3), 571-610.

Suddaby, R., & Greenwood, R. (2005). Rhetorical strategies of legitimacy. *Administrative Science Quarterly, 50*(1), 35-67.

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitmacy. *Academy of Management Annals*, 451-478.

Tallarita, R. (2023, December 5). AI is Testing the Limits of Corporate Governance. *Available at SSRN 4693045*.

Tang, L., Gallagher, C., & Bie, B. (2015). Corporate Social Responsibility Communication Through Corporate Websites: A Comparison of Leading Corporations in the United States and China. *International Journal of Business Communication, 52*(2), 205-227.

Tegmark, M. (2018). *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Vintage.

The Guardian. (2024, February 17). *Microsoft-backed OpenAI valued at $80bn after company completes deal*. Retrieved from The Guardian: https://www.theguardian.com/technology/2024/feb/16/microsoft-openai-valuation-artificial-intelligence

Theis, T. N., & Wong, H.-S. P. (2017). The End of Moore's Law: A New Beginning for Information Technology. *Computing in science & engineering, 19*(2), 41-50.

Tienari, J., Vaara, E., & Björkman, I. (2003). Global capitalism meets national spirit: Discourses in media texts on a cross-border acquisition. *Journal of management inquiry, 12*(4), 377-393.

Tricker, B. (2020). *The Evolution of Corporate Governance.* Cambridge: Cambridge University Press.

Turcan, R. V. (2012). International New Venture Legitimation: An Exploratory Study. *Administrative Sciences, 3*(4), 237-265.

Tylecotea, A., & Ramirez, P. (2006). Corporate governance and innovation: The UK compared with the US and 'insider' economies. *Research Policy, 35*, 160-180.

Vaara, E., & Tienari, J. (2008). A Discursive Perspective on Legitimation Strategies in Multinational Corporations. *The Academy of Management Review, 33*(4), 985-993.

Vaara, E., & Tienari, J. (2011). On the Narrative Construction of Multinational Corporations: An Antenarrative Analysis of Legitimation and Resistance in a Cross-Border Merger. *Organization Science, 22*(2), 370-390.

Vaara, E., Kleymann, B., & Seristö, H. (2003). Strategies as Discursive Constructions: The Case of Airline Alliances. *Journal of Management Studies*, 1-35.

van Dijk, T. A. (1988). *News as Discourse* (1st ed.). Oxon: Routledge.

van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse and Communication, 1*(1), 91-112.

van Leeuwen, T. (2008). *Discourse and Practice: New Tools for Critical Discourse Analysis.* Oxford: Oxford University Press.

van Leeuwen, T. (2013). The Representation of Social Actors. In C. R. Caldas-Coulthard, & M. Coulthard, *Texts and Practices: Readings in Critical Discourse Analysis* (pp. 32-70). Oxfordshire: Taylor Francis.

van Leeuwen, T., & Wodak, R. (1999). Legitimizing Immigration Control: A Discourse-Historical Analysis. *Discourse Studies*, 5-128.

Vaughan, D. (1990). Autonomy, Interdependence, and Social Control: NASA and the Space Shuttle Challenger. *Administrative Science Quarterly, 35*(2), 225-257.

Velte, P. (2023). The link between corporate governance and corporate. *Management Review Quarterly, 73*, 353-411.

Viader, A. M., & Espina, M. I. (2014). Are not-for-profits learning from for-profit-organizations? A look into governance. *Corporate Governance, 14*(1), 1-14.

Waddingham, J., Zachary, M., & Ketchen Jr, D. (2020). Insights on the go: Leveraging business podcasts to enhance organizational performance. *Business Horizons, 63*(3), 275-285.

Wang, J., & Dewhirst, H. (1992). Boards of Directors and Stakeholder Orientations. *Journal of Business Ethics, 11*(1), 115-123.

Wasserman, N. (2006). Stewards, Agents, and the Founder Discount: Executive Compensation in New Ventures. *Academy of Management Journal, 49*(5), 960-976.

Widdowson, H. G. (1995). Discourse Analysis: A Critical View. *Language and Literature, 4*(3), 157-172.

Wright, M., Siegel, D. S., Keasey, K., & Filatotchev, I. (2013). *The Oxford Handbook of Corporate Governance.* Oxford: Oxford University Press.

Yampolskiy, R. (2013). What to Do with the Singularity Paradox? In V. Müller, *Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics* (pp. 397-413). Berlin: Springer.

Zaman, R., Jain, T., Samara, G., & Jamali, D. (2022). Corporate Governance Meets Corporate Social Responsibility: Mapping the Interface. *Business & Society*, 690-752.

Zimmerman, M. A., & Zeitz, G. J. (2002). Beyond Survival: Achieving New Venture Growth by Building Legitimacy. *The Academy of Management Review, 27*(3), 414-431.

Appendix

Appendix 1:

Text 1, OpenAI (OpenAI, 2024)

Our structure

We designed OpenAI's structure—a partnership between our original Nonprofit and a new capped profit arm—as a chassis for OpenAI's mission: to build artificial general intelligence (AGI) that is safe and benefits all of humanity.

We announced our "capped profit" structure in 2019, about three years after founding the original OpenAI Nonprofit.

Since the beginning, we have believed that powerful AI, culminating in AGI—meaning a highly autonomous system that outperforms humans at most economically valuable work—has the potential to reshape society and bring tremendous benefits, along with risks that must be safely addressed. The increasing capabilities of present day systems mean it's more important than ever for OpenAI and other AI companies to share the principles, economic mechanisms, and governance models that are core to our respective missions and operations.

Overview

We founded the OpenAI Nonprofit in late 2015 with the goal of building safe and beneficial artificial general intelligence for the benefit of humanity. A project like this might previously have been the provenance of one or multiple governments—a humanity-scale endeavor pursuing

broad benefit for humankind.

Seeing no clear path in the public sector, and given the success of other ambitious projects in private industry (e.g., SpaceX, Cruise, and others), we decided to pursue this project through private means bound by strong commitments to the public good. We initially believed a 501(c)(3) would be the most effective vehicle to direct the development of safe and broadly beneficial AGI while remaining unencumbered by profit incentives. We committed to publishing our research and data in cases where we felt it was safe to do so and would benefit the public.

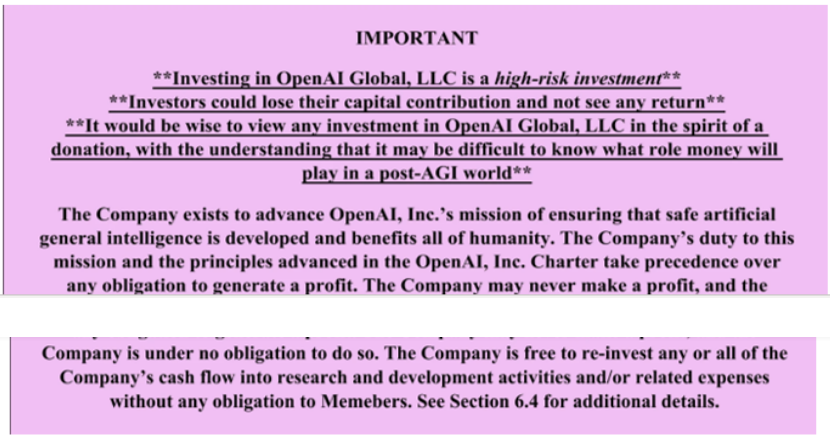
We always suspected that our project would be capital intensive, which is why we launched with the goal of $1 billion in donation commitments. Yet over the years, OpenAI's Nonprofit received approximately $130.5 million in total donations, which funded the Nonprofit's operations and its initial exploratory work in deep learning, safety, and alignment.

It became increasingly clear that donations alone would not scale with the cost of computational power and talent required to push core research forward, jeopardizing our mission. So we devised a structure to preserve our Nonprofit's core mission, governance, and oversight while enabling us to raise the capital for our mission:

- The OpenAI Nonprofit would remain intact, with its board continuing as the overall governing body for all OpenAI activities.
- A new for-profit subsidiary would be formed, capable of issuing equity to raise capital and hire world class talent, but still at the direction of the Nonprofit. Employees working on for-profit initiatives were transitioned over to the new subsidiary.
- The for-profit would be legally bound to pursue the Nonprofit's mission, and carry out that mission by engaging in research, development, commercialization and other core operations. Throughout, OpenAI's guiding principles of safety and broad benefit would be central to its approach.
- The for-profit's equity structure would have caps that limit the maximum financial returns to investors and employees to incentivize them to research, develop, and deploy AGI in a way that balances commerciality with safety and sustainability, rather than focusing on pure profit-maximization.
- The Nonprofit would govern and oversee all such activities through its board in addition to its own operations. It would also continue to undertake a wide range of charitable initiatives, such as sponsoring a comprehensive basic income study (opens in a new window) supporting economic impact research, and experimenting with education-centered programs like OpenAI Scholars. Over the years, the Nonprofit also supported a number of other public charities focused on technology, economic impact and justice, including the Stanford University Artificial Intelligence Index Fund, Black Girls Code, and the ACLU Foundation.

In that way, the Nonprofit would remain central to our structure and control the development of

AGI, and the for-profit would be tasked with marshalling the resources to achieve this while remaining duty-bound to pursue OpenAI's core mission. The primacy of the mission above all is encoded in the operating agreement of the for-profit, which every investor and employee is subject to:



IMPORTANT

****Investing in OpenAI Global, LLC is a *high-risk investment*****
****Investors could lose their capital contribution and not see any return****
****It would be wise to view any investment in OpenAI Global, LLC in the spirit of a donation, with the understanding that it may be difficult to know what role money will play in a post-AGI world****

The Company exists to advance OpenAI, Inc.'s mission of ensuring that safe artificial general intelligence is developed and benefits all of humanity. The Company's duty to this mission and the principles advanced in the OpenAI, Inc. Charter take precedence over any obligation to generate a profit. The Company may never make a profit, and the Company is under no obligation to do so. The Company is free to re-invest any or all of the Company's cash flow into research and development activities and/or related expenses without any obligation to Memebers. See Section 6.4 for additional details.

The structure in more detail

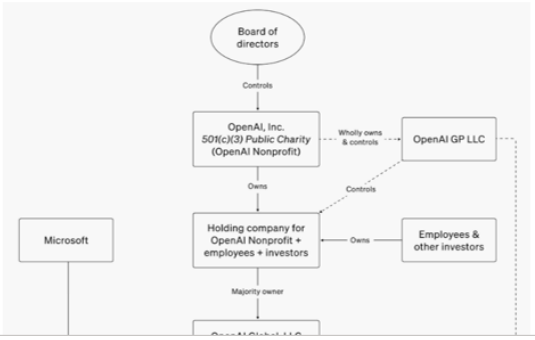
While investors typically seek financial returns, we saw a path to aligning their motives with our mission. We achieved this innovation with a few key economic and governance provisions:

- First, the for-profit subsidiary is fully controlled by the OpenAI Nonprofit. We enacted this by having the Nonprofit wholly own and control a manager entity (OpenAI GP LLC) that has the power to control and govern the for-profit subsidiary.
- Second, because the board is still the board of a Nonprofit, each director must perform their fiduciary duties in furtherance of its mission—safe AGI that is broadly beneficial. While the for-profit subsidiary is permitted to make and distribute profit, it is subject to this mission. The Nonprofit's principal beneficiary is humanity, not OpenAI investors.
- Third, the board remains majority independent. Independent directors do not hold equity in OpenAI. Even OpenAI's CEO, Sam Altman, does not hold equity directly. His only interest is indirectly through a Y Combinator investment fund that made a small investment in OpenAI before he was full-time.
- Fourth, profit allocated to investors and employees, including Microsoft, is capped. All residual value created above and beyond the cap will be returned to the Nonprofit for

the benefit of humanity.

- Fifth, the board determines when we've attained AGI. S Again, by AGI we mean a highly autonomous system that outperforms humans at most economically valuable work.

Such a system is excluded from IP licenses and other commercial terms with Microsoft, which only apply to pre-AGI technology.



We strive to preserve these core governance and economic components of our structure when exploring opportunities to accelerate our work. Indeed, given the path to AGI is uncertain, our structure is designed to be adaptable—we believe this is a feature, not a bug

Microsoft

Shortly after announcing the OpenAI capped profit structure (and our initial round of funding) in 2019, we entered into a strategic partnership with Microsoft. We subsequently extended our partnership, expanding both Microsoft's total investment as well as the scale and breadth of our commercial and supercomputing collaborations.

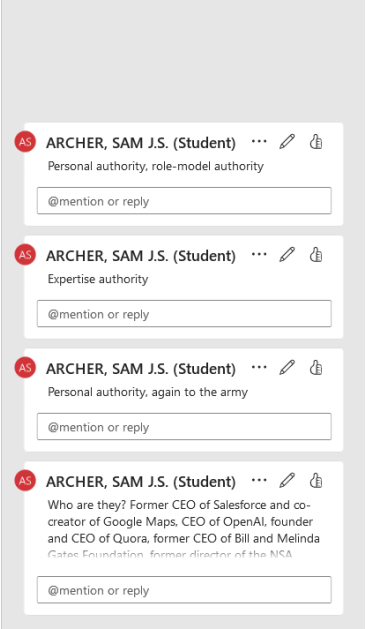
While our partnership with Microsoft includes a multibillion dollar investment, OpenAI remains an entirely independent company governed by the OpenAI Nonprofit. Microsoft is a non-voting board observer and has no control. And, as explained above, AGI is explicitly carved out of all commercial and IP licensing agreements.

These arrangements exemplify why we chose Microsoft as our compute and commercial partner. From the beginning, they accepted our capped equity offer and our request to leave AGI technologies and governance for the Nonprofit and the rest of humanity. They have also worked with us to create and refine our joint safety board that reviews our systems before they are deployed. Harkening back to our origins, they understand that this is a unique and ambitious project that requires resources at the scale of the public sector, as well as the very same

106

107 Our board

108 OpenAI is governed by the board of the OpenAI Nonprofit, currently comprised of
109 Independent Directors Bret Taylor (Chair), Sam Altman, Adam D'Angelo, DJ, Sue
110 Desmond-Hellmann, Retired U.S. Army General Paul M. Nakasone, Nicole Seligman,
111 Fidji Simo, and Larry Summers.



Appendix 2:

(Anthropic, 2023)

The Long-Term Benefit Trust

Today we are sharing more details about our new governance structure called the Long-Term Benefit Trust (LTBT), which we have been developing since the birth of Anthropic. The LTBT is our attempt to fine-tune our corporate governance to address the unique challenges and long-term opportunities we believe transformative AI will present.

The Trust is an independent body of five financially disinterested members with an authority to select and remove a portion of our Board that will grow over time (ultimately, a majority of our Board). Paired with our Public Benefit Corporation status, the LTBT helps to align our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity.

Corporate Governance Basics

A corporation is overseen by its board of directors. The board selects and oversees the leadership team (especially the CEO), who in turn hire and manage the employees. The default corporate governance setup makes directors accountable to the stockholders in several ways. For example:

- Directors are elected by, and may be removed by stockholders.
- Directors are legally accountable to stockholders for fulfilling their fiduciary duties.
- Directors are often paid in shares of stock of the corporation, which helps to align their incentives with the financial interests of stockholders.

Importantly, the rights to elect, remove, and sue directors belong exclusively to the stockholders. Some wonder, therefore, whether directors of a corporation are permitted to optimize for stakeholders beyond the corporation's stockholders, such as customers and the general public. This question is the subject of a rich debate, which we won't delve into here. For present purposes, it is enough to observe that all the key mechanisms of accountability in corporate law push directors to prioritize the financial interests of stockholders.

Fine-tuning Anthropic's Corporate Governance

Corporate governance has seen centuries of legal precedent and iteration, and views differ greatly on its effectiveness, strengths, and weaknesses. At Anthropic, our perspective is that the capacity of corporate governance to produce socially beneficial outcomes depends strongly on

non-market externalities. Externalities are a type of market failure that occurs when a transaction between two parties imposes costs or benefits on a third party who has not consented to the transaction. Common examples of costs include pollution from factories, systemic financial risk from banks, and national security risks from weapons manufacturers. Examples of positive spillover effects include the societal benefits of education that reach beyond the individuals being educated, or investments in R&D that boost entire sectors beyond the company making the investment. Many parties who contract with a corporation, such as customers, workers, and suppliers, are capable of negotiating or demanding prices and terms that reflect the full costs and benefits of their exchanges. But other parties, such as the general public, don't directly contract with a corporation and therefore do not have a means to charge or pay for the costs and benefits they experience.

The greater the externalities, the less we expect corporate governance defaults to serve the interests of non-contracting parties such as the general public. We believe AI may create unprecedentedly large externalities, ranging from national security risks, to large-scale economic disruption, to fundamental threats to humanity, to enormous benefits to human safety and health. The technology is advancing so rapidly that the laws and social norms that constrain other high-externality corporate activities have yet to catch up with AI; this has led us to invest in fine-tuning Anthropic's governance to meet the challenge ahead of us.

To be clear, for most of the day-to-day decisions Anthropic makes, public benefit is not at odds with commercial success or stockholder returns, and if anything our experience has shown that the two are often strongly synergistic: our ability to do effective safety research depends on building frontier models (the resources for which are greatly aided by commercial success), and our ability to foster a "race to the top" depends on being a viable company in the ecosystem in both a technical sense and a commercial sense. We do not expect the LTBT to intervene in these day-to-day decisions or in our ordinary commercial strategy.

Rather, the need for fine-tuning of the governance structure ultimately derives from the potential for extreme events and the need to handle them with humanity's interests in mind, and we expect the LTBT to primarily concern itself with these long-range issues. For example, the LTBT can ensure that the organizational leadership is incentivized to carefully evaluate future models for catastrophic risks or ensure they have nation-state level security, rather than prioritizing being the first to market above all other objecttives.

Baseline: Public Benefit Corporation

One governance feature we have already shared is that Anthropic is a Delaware Public Benefit Corporation, or PBC. Like most large companies in the United States, Anthropic is incorporated in Delaware, and Delaware corporate law expressly permits the directors of a PBC to balance the financial interests of the stockholders with the public benefit purpose specified in the corporation's certificate of incorporation, and the best interests of those materially affected by the corporation's conduct. The public benefit purpose stated in Anthropic's certificate is the responsible development and maintenance of advanced AI for the long-term benefit of humanity. This gives our board the legal latitude to weigh long- and short-term externalities of decisions–whether to deploy a particular AI system, for example–alongside the financial interests of our stockholders.

The legal latitude afforded by our PBC structure is important in aligning Anthropic's governance with our public benefit mission. But we didn't feel it was enough for the governance challenges we foresee in the development of transformative AI. Although the PBC form makes it legally permissible for directors to balance public interests with the maximization of stockholder value, it does not make the directors of the corporation directly accountable to other stakeholders or align their incentives with the interests of the general public. We set out to design a structure that would supply our directors with the requisite accountability and incentives to appropriately balance the financial interests of our stockholders and our public benefit purpose at key junctures where we expect the consequences of our decisions to reach far beyond Anthropic.

LTBT: Basic Structure and Features

The Anthropic Long-Term Benefit Trust (LTBT, or Trust) is an independent body comprising five Trustees with backgrounds and expertise in AI safety, national security, public policy, and social enterprise. The Trust's arrangements are designed to insulate the Trustees from financial interest in Anthropic and to grant them sufficient independence to balance the interests of the public alongside the interests of Anthropic's stockholders.

At the close of our Series C, we amended our corporate charter to create a new class of stock (Class T) held exclusively by the Trust. The Class T stock grants the Trust the authority to elect and remove a number of Anthropic's board members that will phase in according to time- and funding-based milestones; in any event, the Trust will elect a majority of the board within 4

years. At the same time, we created a new director seat that will be elected by the Series C and subsequent investors to ensure that our investors' perspectives will be directly represented on the board into the future.

The Class T stock also includes "protective provisions" that require the Trust to receive notice of certain actions that could significantly alter the corporation or its business.

The Trust is organized as a "purpose trust" under the common law of Delaware, with a purpose that is the same as that of Anthropic. The Trust must use its powers to ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic's conduct and our public benefit purpose.

A Different Kind of Stockholder

In establishing the Long-Term Benefit Trust we have, in effect, created a different kind of stockholder in Anthropic. Anthropic will continue to be overseen by its board, which we expect will make the decisions of consequence on the path to transformative AI. In navigating these decisions, a majority of the board will ultimately have accountability to the Trust as well as to stockholders, and will thus have incentives to appropriately balance the public benefit with stockholder interests. Moreover, the board will benefit from the insights of Trustees with deep expertise and experience in areas key to Anthropic's public benefit mission. Together we believe the insights and incentives supplied by the Trust will result in better decision making when the stakes are highest.

The gradual "phase-in" of the LTBT will allow us to course-correct an experimental structure and also reflects a hypothesis that, early in a company's history, it can often function best with streamlined governance and not too many stakeholders; whereas as it becomes more mature and has more profound effects on society, externalities tend to manifest themselves progressively more, making checks and balances more critical.

A Corporate Governance Experiment

The Long-Term Benefit Trust is an experiment. Its design is a considered hypothesis, informed by some of the most accomplished corporate governance scholars and practitioners in the nation, who helped our leadership design and "red team" this structure. We're not yet ready to hold this out as an example to emulate; we are empiricists and want to see how it works.

One of the most difficult design challenges was reconciling the imperative for the Trust structure to be resilient to end runs while the stakes are high with the reality of the Trust's experimental nature. It's important to prevent this arrangement from being easily undone, but it is also rare to get something like this right on the first try. We have therefore designed a process for amendment that carefully balances durability with flexibility. We envision that most adjustments will be made by agreement of the Trustees and Anthropic's Board, or the Trustees and the other stockholders. Owing to the Trust's experimental nature, however, we have also designed a series of "failsafe" provisions that allow changes to the Trust and its powers without the consent of the Trustees if sufficiently large supermajorities of the stockholders agree. The required supermajorities increase as the Trust's power phases in, on the theory that we'll have more experience–and less need for iteration–as time goes on, and the stakes will become higher.

Meet the Initial Trustees

The initial Trustees are:

Jason Matheny: CEO of the RAND Corporation

Kanika Bahl: CEO & President of Evidence Action

Neil Buddy Shah: CEO of the Clinton Health Access Initiative (Chair)

Paul Christiano: Founder of the Alignment Research Center

Zach Robinson: Interim CEO of Effective Ventures US

The Anthropic board chose these initial Trustees after a year-long search and interview process to surface individuals who exhibit thoughtfulness, strong character, and a deep understanding of the risks, benefits, and trajectory of AI and its impacts on society. Trustees serve one-year terms and future Trustees will be elected by a vote of the Trustees. We are honored that this founding group of Trustees chose to accept their places on the Trust, and we believe they will provide invaluable insight and judgment.

[1] An earlier version of the Trust, which was then called the "Long-Term Benefit Committee," was written into our Series A investment documents in 2021, but since the committee was not slated to elect its first director until 2023, we took the intervening time to red-team and improve the legal structure and to carefully consider candidate selection. The current LTBT is the result.

[2] The Trust structure was designed and "red teamed" with immeasurable assistance by John Morley of Yale Law School, David Berger, Amy Simmerman, and other lawyers from Wilson Sonsini, and by Noah Feldman and Seth Berman from Harvard Law School and Ethical Compass Advisors.

Appendix 3:

Text 3, Altman (Fridman, 2024)

Lex Fridman(00:03:02) But the thing you had a sense that you would experience is some kind of power struggle?

Sam Altman(00:03:08) The road to AGI should be a giant power struggle. The world should... Well, not should. I expect that to be the case.

Lex Fridman(00:03:17) And so you have to go through that, like you said, iterate as often as possible in figuring out how to have a board structure, how to have organization, how to have the kind of people that you're working with, how to communicate all that in order to deescalate the power struggle as much as possible.

Sam Altman(00:03:37) Yeah.

Lex Fridman(00:03:37) Pacify it.

Sam Altman(00:03:38) But at this point, it feels like something that was in the past that was really unpleasant and really difficult and painful, but we're back to work and things are so busy and so intense that I don't spend a lot of time thinking about it. There was a time after, there was this fugue state for the month after, maybe 45 days after, that I was just drifting through the days. I was so out of it. I was feeling so down.

Lex Fridman(00:04:17) Just on a personal, psychological level?

Sam Altman(00:04:20) Yeah. Really painful, and hard to have to keep running OpenAI in the middle of that. I just wanted to crawl into a cave and recover for a while. But now it's like we're just back to working on the mission.

Lex Fridman(00:04:38) Well, it's still useful to go back there and reflect on board structures, on power dynamics, on how companies are run, the tension between research and product development and money and all this kind of stuff so that you, who have a very high potential of building AGI, would do so in a slightly more organized, less dramatic way in the future. So there's

value there to go, both the personal psychological aspects of you as a leader, and also just the board structure and all this messy stuff.

Sam Altman(00:05:18) I definitely learned a lot about structure and incentives and what we need out of a board. And I think that it is valuable that this happened now in some sense. I think this is probably not the last high-stress moment of OpenAI, but it was quite a high-stress moment. My company very nearly got destroyed. And we think a lot about many of the other things we've got to get right for AGI, but thinking about how to build a resilient org and how to build a structure that will stand up to a lot of pressure in the world, which I expect more and more as we get closer, I think that's super important.

Lex Fridman(00:06:01) Do you have a sense of how deep and rigorous the deliberation process by the board was? Can you shine some light on just human dynamics involved in situations like this? Was it just a few conversations and all of a sudden it escalates and why don't we fire Sam kind of thing?

Sam Altman(00:06:22) I think the board members are well-meaning people on the whole, and I believe that in stressful situations where people feel time pressure or whatever, people understand and make suboptimal decisions. And I think one of the challenges for OpenAI will be we're going to have to have a board and a team that are good at operating under pressure.

Lex Fridman(00:07:00) Do you think the board had too much power?

Sam Altman(00:07:03) I think boards are supposed to have a lot of power, but one of the things that we did see is in most corporate structures, boards are usually answerable to shareholders.

Sometimes people have super voting shares or whatever. In this case, and I think one of the things with our structure that we maybe should have thought about more than we did is that the board of a nonprofit has, unless you put other rules in place, quite a lot of power. They don't really answer to anyone but themselves. And there's ways in which that's good, but what we'd really like is for the board of OpenAI to answer to the world as a whole, as much as that's a practical thing.

Lex Fridman(00:07:44) So there's a new board announced.

Sam Altman(00:07:46) Yeah.

Lex Fridman(00:07:47) There's I guess a new smaller board at first, and now there's a new final board?

Sam Altman(00:07:53) Not a final board yet. We've added some. We'll add more.

Lex Fridman(00:07:56) Added some. Okay. What is fixed in the new one that was perhaps broken in the previous one?

Sam Altman(00:08:05) The old board got smaller over the course of about a year. It was nine and then it went down to six, and then we couldn't agree on who to add. And the board also I think didn't have a lot of experienced board members, and a lot of the new board members at OpenAI have just have more experience as board members. I think that'll help.

Lex Fridman(00:08:31) It's been criticized, some of the people that are added to the board. I heard a lot of people criticizing the addition of Larry Summers, for example. What's the process of selecting the board? What's involved in that?

Sam Altman(00:08:43) So Brett and Larry were decided in the heat of the moment over this very tense weekend, and that weekend was a real rollercoaster. It was a lot of ups and downs. And we were trying to agree on new board members that both the executive team here and the old board members felt would be reasonable. Larry was actually one of their suggestions, the old board members. Brett, I think I had even previous to that weekend suggested, but he was busy and didn't want to do it, and then we really needed help in [inaudible 00:09:22]. We talked about a lot of other people too, but I felt like if I was going to come back, I needed new board members. I didn't think I could work with the old board again in the same configuration, although we then decided, and I'm grateful that Adam would stay, but we considered various configurations, decided we wanted to get to a board of three and had to find two new board members over the course of a short period of time.

(00:09:57) So those were decided honestly without... You do that on the battlefield. You don't have time to design a rigorous process then. For new board members since, and new board members we'll add going forward, we have some criteria that we think are important for the board to have, different expertise that we want the board to have. Unlike hiring an executive where you need them

to do one role well, the board needs to do a whole role of governance and thoughtfulness well, and so, one thing that Brett says which I really like is that we want to hire board members in slates, not as individuals one at a time. And thinking about a group of people that will bring nonprofit expertise, expertise at running companies, good legal and governance expertise, that's what we've tried to optimize for.

Lex Fridman(00:10:49) So is technical savvy important for the individual board members?

Sam Altman(00:10:52) Not for every board member, but for certainly some you need that. That's part of what the board needs to do.

Lex Fridman(00:10:57) The interesting thing that people probably don't understand about OpenAI, I certainly don't, is all the details of running the business. When they think about the board, given the drama, they think about you. They think about if you reach AGI or you reach some of these incredibly impactful products and you build them and deploy them, what's the conversation with the board like? And they think, all right, what's the right squad to have in that kind of situation to deliberate?

Sam Altman(00:11:25) Look, I think you definitely need some technical experts there. And then you need some people who are like, "How can we deploy this in a way that will help people in the world the most?" And people who have a very different perspective. I think a mistake that you or I might make is to think that only the technical understanding matters, and that's definitely part of the conversation you want that board to have, but there's a lot more about how that's going to just impact society and people's lives that you really want represented in there too.

Appendix 4:

Text 4, Amodei (Patel, 2023)

(01:47:01) - Anthropic's Long Term Benefit Trust

Dario Amodei (01:47:01 - 01:47:49):

Even us who have not been super focused on commercialization and more on safety, the graph goes up and it goes up relatively quickly. I can only imagine what's happening at the orgs where this is their singular focus. It's certainly happening fast but it's an exponential from the small base while the technology itself is moving fast.

It's a race between how fast the technology is getting better and how fast it's integrated into the economy. And I think that's just a very unstable and turbulent process. Both things are going to happen fast but if you ask me exactly how it's going to play out, exactly what order things are going to happen, I don't know. And I'm skeptical of the ability to predict.

Dwarkesh Patel (01:47:49 - 01:48:14):

I'm curious. With regards to Anthropic specifically, you're a public benefit corporation and rightfully so, you want to make sure that this is an important technology. Obviously, the only thing you want to care about is not shareholder value.

But how do you talk to investors who are putting in hundreds of millions, billions of dollars of money? How do you get them to put in this amount of money without the shareholder value being the main concern?

Dario Amodei (01:48:14 - 01:49:18):

I think the LTBT (Long Term Benefit Trust) is the right thing on this. We're going to talk more about the LTBT, but some version of that has been in development since the beginning of Anthropic, even formally. Even as the body has changed, from the beginning, it was like, this body is going to exist and it's unusual.

Every traditional investor who invests in Anthropic looks at this. Some of them are just like, whatever, you run your company how you want. Some of them are like, oh my god, this body of random people could move Anthropic in a direction that's totally contrary to shareholder value. Now there are legal limits on that, of course, but we have to have this conversation with every investor. And then it gets into a conversation of, well, what are the kinds of things that we might do that would be contrary to the interests of traditional investors. And just having those conversations has helped get everyone on the same page.

Dwarkesh Patel (01:49:18 - 01:49:43):

I want to talk about the fact that so many of the founders and the employees at Anthropic are physicists. We talked in the beginning about the scaling laws and how the power laws from physics are something you see here, but what are the actual approaches and ways of thinking from physics that seem to have carried over so well? Is that notion of effective theory super useful? What is going on here?

Dario Amodei (01:49:43 - 01:50:18):

Part of it is just that physicists learn things really fast. We have generally found that if we hire someone who is a Physics PhD or something, that they can learn ML and contribute just very quickly in most cases. And because several of our founders myself, Jared Kaplan, Sam McCandlish were physicists, we knew a lot of other physicists, and so we were able to hire them. And now there might be 30 or 40 of them here. ML is not still not yet a field that has an enormous amount of depth, and so they've been able to get up to speed very quickly.

Dwarkesh Patel (01:50:18 - 01:50:41):

Are you concerned that there's a lot of people who would have been doing physics or something, they would've gone into finance instead and since Anthropic exists, they have now been recruited to go into AI. You obviously care about AI safety, but maybe in the future they leave and they get funded to do their own thing. Is that a concern that you're bringing more people into the ecosystem here?

Dario Amodei (01:50:41 - 01:51:18):

There's a broad set of actions, like we're causing GPUs to exist. There's a lot of side effects that you can't currently control or that you just incur if you buy into the idea that you need to build frontier models. And that's one of them. A lot of them would have happened anyway. I mean, finance was a hot thing 20 years ago, so physicists were doing it. Now ML is a hot thing, and it's not like we've caused them to do it when they had no interest previously. But again, at the margin, you're bidding things up, and a lot of that would have happened anyway. Some of it wouldn't but it's all part of the calculus.

[...]

Dwarkesh Patel (01:56:14 - 01:56:26):

You've been less public than the CEOs of other AI companies. You're not posting on Twitter, you're not doing a lot of podcasts except for this one. What gives? Why are you off the radar?

Dario Amodei (01:56:26 - 01:58:03):

I aspire to this and I'm proud of this. If people think of me as boring and low profile, this is actually kind of what I want. I've just seen cases with a number of people I've worked with, where attaching your incentives very strongly to the approval or cheering of a crowd can destroy your mind, and in some cases, it can destroy your soul.

I've deliberately tried to be a little bit low profile because I want to defend my ability to think about things intellectually in a way that's different from other people and isn't tinged by the approval of other people. I've seen cases of folks who are deep learning skeptics, and they become known as deep learning skeptics on Twitter. And then even as it starts to become clear to me, they've sort of changed their mind. This is their thing on Twitter, and they can't change their Twitter persona and so forth and so on.

I don't really like the trend of personalizing companies. The whole cage match between CEOs approach. I think it distracts people from the actual merits and concerns of the company in question. I want people to think in terms of the nameless, bureaucratic institution and its incentives more than they think in terms of me. Everyone wants a friendly face, but actually, friendly faces can be misleading.