Media@LSE MSc Dissertation Series

Editors: Simidele Dosekun and Hao Wang

CATASTROPHIC YET BENEFICIAL

A Critical Discourse Analysis of OpenAI's Narratives on Existential Threats and Present Harms

CAMILA MOLINA AVILA

Published by Media@LSE, London School of Economics and Political Science ("LSE"), Houghton Street, London WC2A 2AE. The LSE is a School of the University of London. It is a Charity and is incorporated in England as a company limited by guarantee under the Companies Act (Reg number 70527).

ABSTRACT

This dissertation critically examines OpenAI's public discourse on AI risks and safety, exploring how the company's communication strategies emphasize future existential threats while potentially obscuring current real harms. Through a detailed analysis of OpenAI's narratives, the research investigates the implications of this discourse for societal understanding and policy-making. Using Wodak's Discourse-Historical Approach (DHA) as the methodological framework, the study reveals how OpenAI constructs a dual narrative of AI as both a solution to humanity's greatest challenges and a source of catastrophic risks. The findings suggest that by focusing on speculative future scenarios, OpenAI may divert attention from ongoing issues such as bias, disinformation, and economic disruption caused by current AI technologies. This strategic discursive emphasis on future existential risks could lead to a regulatory system that prioritizes future risk mitigation over the immediate regulation of present-day AI applications.

INTRODUCTION

The rapid advancement of artificial intelligence (AI) has sparked intense global debate, not only about the transformative potential of these technologies but also about the profound risks they carry. Among the leading voices in this discourse is OpenAI, a company that has positioned itself at the forefront of AI development and safety. With its mission to ensure that 'Artificial General Intelligence (AGI) benefits all of humanity'[1], OpenAI's public communications have become a powerful influence on how society perceives the promises and dangers of AI.

This dissertation interrogates the narratives and strategies employed by OpenAI to shape public and policy perceptions of AI. It critically examines how OpenAI's discourse often fluctuates between utopian visions of AI solving humanity's greatest challenges and dystopian warnings of AI-driven existential threats. By focusing predominantly on speculative future scenarios, OpenAI may be diverting attention from the very real, present and ongoing harms associated with AI, such as bias, disinformation, economic inequality, labor exploitation, among others. This narrative choice has significant implications, potentially skewing societal understanding and influencing policy-making in ways that prioritize long-term existential risk mitigation over urgent and necessary regulation of current AI applications.

As this dissertation describes, it becomes evident that OpenAI's discourse is not merely a reflection of its ethical considerations but a strategic construction aimed at maintaining its leadership in the AI field while shaping the global narrative around AI risks. Through a detailed analysis grounded in Wodak's Discourse-Historical Approach (DHA), this work explores how OpenAI crafts its public image, balances its responsibilities, and navigates the complex ethical landscape of AI development. The findings raise critical questions about the role of tech companies in framing societal debates and the potential consequences of allowing these narratives to overshadow the immediate challenges posed by AI technologies.

In an era where AI's impact is rapidly expanding, the need to scrutinize the discourses shaping its development has never been more urgent. This dissertation aims to contribute to this vital

[1] Full OpenAI charter: https://openai.com/charter/

conversation by revealing the underlying dynamics of OpenAI's public communications and their broader implications for society and policy.

ABOUT OPENAI

The following information about OpenAI is the result of extensive research from diverse sources, including media articles, the OpenAI webpage, podcasts, and videos, to provide historical context that enables a critical examination to effectively answer the research question.

Profile

OpenAI was founded in December 2015 by a group of men in the tech industry, including Elon Musk, Sam Altman, Greg Brockman, Ilya Sutskever, John Schulman, and Wojciech Zaremba. First, established as a non-profit research organization, OpenAI's initial mission was to ensure that artificial general intelligence (AGI) benefits all of humanity. This commitment was underpinned by a guiding principle of openness, which was reflected in the organization's name and its early pledges to share research findings and collaborate with the broader AI community.

The formation of OpenAI was motivated by these men growing recognition that AI could become a transformative and potentially disruptive force in society. The founders expressed their concerns about the risks associated with powerful AI systems being in hands of one single company, presumably Google. From the outset, OpenAI positioned itself as a proactive leader in the responsible development of AI, striving to align the technology's progress with ethical considerations and societal good.

Early Years

In its early years, OpenAI gained attention for its ambitious goals and its commitment to transparency. The organization released several influential research papers, developed notable basic AI models, and contributed to the open-source community by releasing tools and datasets for AI training. This period was marked by a clear emphasis on collaboration and sharing, as OpenAI sought to build trust and position itself as a leading advocate for the ethical development of AI.

Transition to OpenAI LP

In 2019, OpenAI underwent a significant organizational transformation, restructuring itself as a for-profit company known as OpenAI LP, with a 'capped-profit' model. According to OpenAI, this structure was designed to attract the necessary capital and talent to achieve AGI, while maintaining a commitment to the original mission of benefiting humanity. The shift to a for-profit entity was met with criticism and skepticism, as it appeared to conflict with OpenAI's founding principles of openness and transparency.

Also, this transition marked a turning point in OpenAI's public communications. While the organization continued to emphasize its mission-driven focus, its messaging began to reflect a more complex balancing act between advancing AI technology and managing the associated risks. OpenAI's public statements increasingly highlighted the existential threats posed by AGI, alongside the potential for AI to address global challenges. This dual narrative served to reinforce OpenAI's positioning as both a leader in AI innovation and a responsible actor of the technology's development.

Strategic Partnerships and Technological Breakthroughs

As OpenAI's research and capabilities advanced, the organization formed strategic partnerships to further its goals. One of the most significant of these partnerships was with Microsoft, which invested $1 billion in OpenAI in 2019 and became its cloud provider and partner, acquiring 49% of the company but with no direct influence over its board of directors. This partnership facilitated the development and scaling of OpenAI's most advanced models and led towards the launched of ChatGPT 3.5 in November of 2022, a chatbot based on a large language model (LLM) which was able to fluently answer users' inquiries on almost any subject. More models with increasing capabilities have been launch since then.

OpenAI's technological breakthroughs, such as the GPT series, and DALL-E, its text to image generator, have not only demonstrated the power of AI tools but also materialize the risks associated with deploying these technologies at scale. In response, the organization's discourse began to focus more heavily on the importance of aligning AI systems with human values and mitigating the risks of misuse. This narrative shift reflects OpenAI's growing influence in shaping public and policy

discussions around AI, as it navigates the tension between promoting AI's benefits and addressing the complex challenges it poses.

LITERATURE REVIEW

The following section presents a comprehensive literature review on the implications of technology-focused narratives for societal understanding and policy-making, examining key theories, studies, and debates that shape the current understanding of this subject. The insights gained from this analysis provide a strong foundation for the present project, guiding its research framework and methodology.

The Ideological Role of Technology Discourse

Language and technology have long been related, not just through a technical dimension, but through a much more nuanced aspect that characterizes the ways in which both have the power to shape human imagination, beliefs and interactions. This nuanced layer reflects the dialectical relationship of how technology influences the ways we communicate, and also how language, in turn, shapes technology's role in our societies. This paradigm is not new, but it can be difficult to grasp, partly because we have been constantly exposed to an institutionalized dominant discourse about technology, and also because we have experienced the benefits ourselves, or at least we are made to think we have.

The existing literature on the power of technology discourse is abundant, and scholars from diverse fields have been discussing the subject for a very long time. For example, Mosco (2004) focuses on how myths about technology are used to depoliticize socio-political debates, thereby consolidating existing power structures and making other solutions seem impractical or impossible. The myths that, through repetition, turn into beliefs and ideologies have several repercussions, such as the naturalization of the idea of technology as inevitable and the legitimation of certain practices in the name of innovation and development.

One very established myth about technology is that it is neutral and does not contain human bias because it is only numbers processed by a machine, so it becomes the fairest solution for critical decision-making scenarios like, for example, access social-welfare programs, loans, health and education. Morozov (2013) defines this way of thinking as 'tech solutionism', where usually powerful

hegemonic institutions—tech companies and governments—reframe 'all complex social situations either as neatly defined problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized.'

Closely linked to Morozov's view is the technological determinist argument, which states that the 'development of digital technology […] will empower people out of radical inequalities, while naturalizing market-based solutions to every issue of governance' (Brevini, 2021). Techno-determinism is based on a political-economic standpoint that believes technology has the right tools to fix our current broken capitalist systems and will lead us to a future where there is no economic and social crisis because everything is optimized by technology.

One prominent critique of the limitless transformative power of technology is an article called 'The Californian Ideology', published in 1995 by the scholars Cameron and Barbrook. In the article, the authors critique the emerging tech industry in Silicon Valley and its ideology, which merges a libertarian belief in individual freedom and minimal government interference with a utopian vision that sees the internet and technology as inherently liberating forces that will create a decentralized, meritocratic society where investors, entrepreneurs, and private companies would take the place of traditional power structures such as governments, states, and civil society. For the authors, the Californian Ideology's core function is

> *to legitimize the position of the 'virtual class' of the San Francisco Bay Area by portraying them as those who will ensure future prosperity, bolstering their position not just in the United States, but internationally (Hepp, Schmitz and Schneider, 2023).*

Although the essay was published in 1995, the arguments remain relevant more than 25 years later. In a recent paper, Creech and Maddox (2024) examine how tech power has become naturalized in media discourses. The paper specifically focuses on Facebook's CEO, Mark Zuckerberg, and analyzes how he has become central to discussions around the moral responsibilities, regulatory challenges, and political-economic power of tech companies. One of the main findings of the paper is, although critiques and legitimizations are present in media narratives, they both contribute to reinforcing the idea that the current tech power dynamics are natural and beyond the reach of political or societal intervention.

On the basis of studies like this one, it becomes possible to visualize how the way society talks about technology—including CEOs, products, the industry, and more—is deeply connected to our understanding, expectations, interactions, and even counteractions with it. A practical example of this is the use of the word 'cloud´ to refer to a network of remote servers used to store, manage, and process data. Here, the term ´cloud' is used as a metaphorical device to represent a complex infrastructure as an intangible object, highlighting its characteristic of ubiquity. Although a word may seem harmless, in line with Bourdieu's theory of symbolic power, this has significant ramifications in our perception of what technology is, how it works and what are its benefits and disadvantages. In this case, the word choice is obscuring the environmental damage caused by the tangible infrastructures 'the cloud' needs to operate, as well as the political implications of managing private personal data.

In the field of technology, the use of metaphors to describe its capabilities is a very common strategy. In a short essay, Wyatt (2021) explores how metaphors have historically been used by academics, tech companies, and regulators to describe the Internet and its possibilities with words like frontier, highway, and library. Wyatt (2021) also pays attention to more current metaphors that are framed around nature, such as 'data as oil' or 'data flows', which suggest that data is a resource to be extracted and managed. This conclusion aligns with Couldry and Mejias's (2019) main argument in their data colonialism theory. Under this lens, metaphors can be seen as effective rhetorical devices strategically used by a group of actors to naturalize technological solutions as the only correct way to address problems and to depoliticize their effects on society. This is achieved by reinforcing the myth of neutrality and positioning the field as a purely technical player rather than a socio-political one.

Similarly, the field of AI is no exception to the trend of anthropomorphization[2], a discursive strategy that is not exclusive to technology. This strategy has been extensively employed by tech industry experts to describe AI capabilities and has been echoed by other powerful actors such as universities, governments, media companies, and the creative industry. Rehak (2021) highlights how the persistent use of anthropomorphized terms to describe AI technologies can effectively blur the line between what is technically feasible and what is imagined. Rehak illustrates this with the

[2] According to the Merriam-Webster Dictionary anthropomorphize means to attribute human form or personality to things not human.

example of artificial neural networks[3], which are frequently explained and described in terms of the functions of the human brain. This rhetorical practice

> *opens up the metaphorical space to other neighboring, yet misleading, concepts. For instance, scientists often do not speak of networks being 'configured' but rather of them being 'trained' or engaging in '(deep) learning'. Related notions include 'recognition', 'acting', 'discrimination', 'communication', 'memory', 'understanding', and, of course, 'intelligence' (Rehak, 2021: 93).*

The consequences of linking AI technology to human actions are at least three and are interconnected. First, there is the risk that these terms will be interpreted too literally by the general public, leading to misguided expectations about AI's capabilities. Second, the sense of agency generated by this practice can create a virtual separation between AI models and their designers and owners, resulting in a grey area where accountability measures become difficult to enforce when problems arise. Third, by misrepresenting technical capabilities and implicitly attributing autonomous agency to technology through specific language and discourse, tech companies gain the power to control the public narrative of AI, shaping and framing how people perceive AI's benefits and risks.

The aim of this sections was to explore the entrenched and powerful relation between discourse and technology as it is an essential element of the main argument of this study. The literature on the discourse of technology reveals the profound influence that language and rhetoric have on shaping our understanding of technology and its role in society. Although there are many more possible topics to touch on, I have focused in the use of metaphors, myths, and anthropomorphized language as they are three macro-strategies that OpenAI uses in its public communications. The narratives constructed through these rhetorical vehicles, not only oversimplify complex and nuanced technological concepts but also serve to reinforce specific ideologies, often aligning with the interests of powerful institutions.

The ideological role of technology discourse has historically contributed to shape public perceptions, influence regulations, naturalize particular agendas, social orders, and obscured the socio-political implications of the technology's deployment. In this framework, the critical examination of these

[3] Rehak (2021) defines artificial neural networks as an approach of computer science to solve complex problems that are hard to explicitly formulate, or more concretely to program (92).

narratives is essential to make visible the subtle ways in which technology discourse contributes to the maintenance of existing power dynamics and the reproduction of inequalities.

Power Dynamics

As stated earlier, the language we use to discuss technology, particularly AI, plays a significant role in shaping our understanding and interaction with it. Therefore, it is crucial to pay attention to the prominent AI narratives that describe society's past, present, and future relationship with technology, as these narratives reveal recurring themes and patterns within different contexts. Therefore, examining AI narratives involves answering questions such as: Which scenarios are being highlighted or ignored? Whose views and values are being privileged? Who or what is being marginalized or entirely excluded from discussions about technology and AI? What are the underlying purposes of these narratives? And finally, how do these factors contribute to the broader sociopolitical perception of AI?

Identifying the specific context from which these AI narratives originate is particularly important, as it reveals the underlying factors that contribute to their success, scrutiny, or invisibility. AI discourse is primarily shaped and dominated by Big Tech companies and, more specifically, by 'a relative handful of mostly male, mostly white and East Asian, mostly young, mostly affluent, highly educated technoscientists and entrepreneurs' (Bones et al., 2021). Perhaps more than in other fields, the hegemonic demographic controlling the tech industry is strikingly clear. It is within this very specific context that public imaginaries of AI technology are formulated.

The concept of imaginaries has been a prominent framework in numerous social science studies exploring the interplay between discourse, technology, and society (i.e. Appadurai, 1996; Taylor, 2004; Marcus, 1995; Flichy, 2007; Mansell, 2012). Imaginaries can generally be defined as collective social visions of the future that emerge as relevant narratives and are mobilized by diverse stakeholders, in turn influencing present practices (Mager and Katzenbach, 2021: 225). Another notable, though sometimes contested, concept is that of sociotechnical imaginaries (SI), developed by Jasanoff and Kim in 2009 in the context of nuclear energy and revised by Jasanoff in 2015. According to Jasanoff, SI are defined as

> *collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology (2015: 10).*

However, more recent studies have concluded that SI are diverse and contested rather than unified visions of the future materialized solely by state actors. The shaping of digital innovations and associated social practices is not driven by governments alone; Big Tech companies, CEOs like Sam Altman of OpenAI, controversial figures like Elon Musk, media outlets, technology journalists, research groups, activists, and civil society organizations all play a role (Mager and Katzenbach, 2021: 226)

This research aligns with the latter perspective. Nonetheless, it is crucial to note that the prominence—and thus the acceptance and widespread nature—of some SI over others is influenced by significant power asymmetries, resource availability, and the public reach and exposure of those who enact them. In recent years, AI industry stakeholders have taken a leading role in the current discussion surrounding this technology, significantly shaping public opinion, political decisions, and media coverage. In these spaces, tech companies have effectively framed AI as both inevitable and essential for addressing major societal challenges (Richter, et al., 2023). Through their public communications and political lobbying campaigns, such as Sam Altman's 2023 global tour to discuss AI with world leaders, companies like OpenAI are promoting their own SI of AI as authoritative voices in the field, while obscuring their economic goals and interests.

Moreover, SI are not static; they are sensitive to time and context, evolving in response to shifting political power, economic stability and crises, private interests, other technological developments, and societal issues. An accurate illustration of this is provided by ten Oever (2021), who examines how the Internet's original SI of egalitarian connectivity, unregulated freedom to innovate, and openness have been reconfigured due to economic and corporate influences, resulting in a more centralized and highly gatekept network. Specifically, ten Oever highlights the role of corporate-driven protocols in altering the Internet's technological affordances, prioritizing corporate interests over the initial vision of equality and innovation, and concludes that the self-regulatory model of governance has failed to sustain the Internet's original egalitarian ideals.

The ideology of 'technology self-regulation in the name of freedom and innovation' can be seen as a significant driver behind the current massive concentration of power and influence wielded by Big

Tech companies. This ideology was embodied in Section 230, described by authors like Jeff Kosseff as 'the twenty-six words that created the internet'. Section 230, a provision of the Communications Decency Act of 1996[4], grants immunity to online platforms by shielding them from liability for user-generated content, allowing these platforms to operate without being classified as publishers or speakers of the content posted by their users (U.S. Department of Justice, 2020). Although the causes of power asymmetries are undoubtedly more complex and nuanced than a single law, this context is essential for understanding how historical ideals and core values are materialized into laws, and how laws initially framed as protections of rights can perpetuate inequalities and deepen the divide between the public and private sectors.

Despite AI being a global issue with widespread effects, it is crucial to recognize that the tech companies shaping Western societies are predominantly based in the United States. Therefore, the political power and influence of the US government must be considered, as the US government's decisions on regulating and framing AI's benefits and risks will impact the rest of the world, particularly countries in the Global South. In conclusion, the control of narratives by powerful private organizations, which obscure their interests under the guise of advocating for broadly beneficial technologies, poses a significant risk of depoliticizing crucial issues. Specifically, current AI narratives have the potential to divert state and civil society actors from engaging in meaningful political debates about AI's present harms by shifting the conversation towards future potential existential risks.

The Utopia and Dystopia of AI

AI is not new technology, as well as the belief -currently more a premonition- that in a near future AI machines will be smarter than humans and that this will completely transform humanity forever. In fact, this has been and still is the base of incredibly successful sci-fi movies and novels all around the world. Also, this belief, according to Ballatore and Natale (2023), 'has raised both utopian hopes and dystopian fears around AI as increased informational automation was identified as a harbinger of unemployment, alienation, surveillance, and excessive bureaucratic control'.

[4] More information about Section 230 of he Communications Decency Act of 1996 can be found here: https://www.eff.org/issues/cda230

Dichotomous narratives have surrounded AI since its emergence in the 1950s, leading to highly contested standpoints about where the technology should be headed, how should be regulated, how should be deployed and what should be the emergency plan for when the dystopian imaginary comes true. In this context,

> *controversies are a constitutive component of the AI myth, as they help to keep it alive and are able to attract attention and space in scientific debate and the public sphere. In fact, […] scientific controversies represent a context through which paradigms, theories, and fields build their influence within the scientific world and, at the same time, in the public and popular arena (Ballatore and Nale, 2023).*

Hence, the use of dichotomous narratives and controversies is not only highly influential in keeping AI relevant on both private and public agendas, but also serves to continually project the conversation into the future. This forward-looking focus often overlooks the harms that the unregulated deployment of existing AI tools has caused in the past and present. Consequently, AI is frequently framed within a narrative of conflict, where its potentially huge benefits are always weighed against its potentially catastrophic risks

Recent research by Ferri and Gloerich (2023) explores the perspectives of present harms and future risks within the AI discourse, analyzing the Future of Life Institute's (FLI) public letter advocating for a six-month moratorium on training big AI models released in March 2023 and signed by a vast number of noticeable AI field personalities and the Distributed Artificial Intelligence Research Institute's (DAIR) response to that letter, published a few days later.

The paper highlights that even though both organizations focus on AI regulation development, their goals are completely different, ultimately reflecting the broader debates on the impact of AI technology in society. The existential risk perspective endorsed by FLI, emphasizes the pressing need for preventing future catastrophic threats posed by the fast passed development to get to superintelligent AI and advocates for robust governance. In contrast, the ongoing harm perspective adopted by DAIR, addresses current, tangible issues such as worker exploitation, bias, lack of transparency, exacerbation of existing inequalities, among other, calling for immediate regulatory measures and corporate accountability (Ferri and Gloerich, 2023).

I consider is important to highlight that demanding accountability and effective regulation for present harms and thinking about the future impacts and consequences of AI are not mutually exclusive

perspectives, and certainly both are necessary and important. Inclusivity and diversity of standpoints are needed in AI public conversations. However, the problem lies on the rhetorical strategies used by Big Tech companies like OpenAI to describe current and future model affordances, that significantly influence how AI risks are assed and understood and what safety measures are prioritized.

One highly influential academic that addresses AI ongoing harms is Emily M. Bender, which co-authored the paper 'The Dangers of Stochastic Parrots: Can Language Models Be Too Big?' (et al., 2021). This groundbreaking paper, critically examines the risks associated with large language models (LLMs) focusing on environmental impact, the reinforcement of harmful biases, and the lack of transparency in AI development processes. The authors argue that the persistent pursuit of ever-larger AI models, driven by a competitive tech industry, overlooks the significant social and ethical concerns and highlight the current ongoing harms that the deployment of such models can perpetuate like existing inequalities, deepen societal divides, reinforce discriminatory practices embedded in the training data and the high environmental cost of training massive AI models (Bender et al., 2021).

When published, this paper sparked a significant debate within the AI community, leading to further discussions about the ethical implications of AI development and the responsibilities of researchers and corporations. This controversy reflects the broader tensions within the AI discourse, where the drive for technological innovation often clashes with the need for ethical and sustainable practices. In this context, this study serves as a critical reminder of the importance of addressing both current and future harms associated with AI.

Finally, one of the pressing issues within the AI discourse is the tendency to hype the affordances of current AI systems, often leading to a skewed understanding of their capabilities and limitations. Big Tech companies, in their pursuit of market dominance, frequently exaggerate the potential benefits and downplay the inherent risks associated with AI technologies. This creates a misleading narrative that these systems are far more advanced and infallible than they actually are. Such hype can lead to a lack of critical scrutiny, delaying necessary regulatory actions and allowing harmful consequences to proliferate unchecked (Hicks, Humphries, and Slater, 2024).

Overhyping AI's capabilities can promote unrealistic expectations among users, policymakers, and the public, which in turn can result in underestimating the need for robust governance frameworks that address ethical, social, and environmental concerns. This misrepresentation also shifts the focus

away from addressing the ongoing harms these technologies perpetuate in favor of an optimistic yet ungrounded vision of AI's future potential.

CONCEPTUAL FRAMEWORK AND RESEARCH QUESTION

This dissertation is grounded in Critical Discourse Analysis (CDA), specifically utilizing Wodak's Discourse-Historical Approach (DHA) to examine how OpenAI's public discourse shapes societal understanding and policy-making around AI risks. Additionally, within a broader theoretical framework, this study adopts a feminist lens alongside a Critical Data Studies approach, as well as perspectives from the field of Media and Communications studies.

The central research question of this study is:

To what extent does OpenAI's public discourse on AI risks obscure current real harms by emphasizing future existential threats, and what are the implications for societal understanding and policy-making?

The rationale for choosing this topic is based on the growing influence of AI and the need for regulatory frameworks that address both present and future challenges. By investigating how OpenAI's discourse may take away attention from current AI harms, this research aims to contribute to a deeper understanding of the relationship between AI development, public perception, and policy. The findings will offer insights into the role of public narratives in shaping AI governance, and potentially informing more balanced and effective regulatory approaches.

RESEARCH DESIGN AND METHODOLOGY

In the following section, I provide a detailed examination of the methods and data utilized in this study to explore the outlined research question.

Methodological Justification

Discourse Studies refer to a range of approaches for studying texts from various theoretical backgrounds. Inside this broad field, one of the most widely adopted approaches is Critical Discourse Analysis (CDA), which aims 'to systematically explore often opaque relationships of causality and

determination between (a) discursive practices, events, and texts, and (b) wider social and cultural structures, relations, and processes' (Fairclough, 1993). In line with its main goal, CDA sees discourse as a form of social practice; as such, it both shapes and is shaped by power relations, ideologies, and institutions. (Fairclough and Wodak, 1997).

The richness of using CDA as a methodology comes from its critical dimension. By adopting a critical stance, it is possible to highlight how our use of language is deeply interconnected with broader underlying goals, traditions, social imaginaries, political institutions, and power dynamics that are not always explicitly obvious or even easy to detect. Hence, there is a pressing need to explore 'how the opacity of these relationships between discourse and society is itself a factor securing power and hegemony' (Fairclough, 1993).

It is important to mention that CDA rejects the notion that language is a neutral tool for describing the world and our interactions. From this perspective, CDA questions what is conceived as naturalized knowledge—those aspects of the world considered 'normal' or 'the way things are'. According to CDA, our personal understanding of the world is always historically and culturally specific, influenced by our experiences, social class, gender, among other factors (Gill, 2011).

As Fairclough and Wodak (1997: 262) mention, there are many theoretical approaches to CDA. While some might take a deeper focus on aspects like the dialogical properties of language, others view the historical context as the main common thread for analysis. Against a CDA theoretical and methodological background, this study adopts the Discourse-Historical Approach (DHA) developed by Ruth Wodak to closely analyse the discursive construction of risks associated with AI in Open AI´s public statements in three different genres (types of data). According to the authors,

> *the distinctive feature of this approach is its attempt to integrate systematically all available background information in the analysis and interpretation of the many layers of a written or spoken text (Fairclough and Wodak, 1997: 266).*

Although DHA was originally designed in the 1990s to analyze implicit discriminatory discourses, one of its most notable characteristics is its interdisciplinarity—combining methods, concepts, and theories from different academic disciplines to address complex research questions. For this study, which comes from the broad discipline of media studies intertwined with science and technology studies and sociology, the emphasis on the historical perspective provides the necessary analytical

tool to identify and examine current discourses (AI risk narratives) that tend to sound 'new' or 'emergent', but that in reality have been used for a very long time by many actors in the tech industry to legitimize certain practices and avoid accountability.

Another important element of DHA is the recognition that background knowledge plays a critical role in the audience's ability to interpret discourses (Fairclough and Wodak, 1997: 266). With technology, it is often the case that those with a technical understanding might recognize legitimate concerns about risks, while others might lack the awareness to see how current harms are being downplayed. In this sense, DHA is the correct method to explore how different stakeholders' understanding of AI risks are shaped by specific discursive strategies and lexical choices.

Ultimately, the DHA approach focuses on examining how implicit discourse often enables 'text producers' to avoid responsibility for their vague utterances. For example, by focusing on future risk narratives, OpenAI could potentially evade responsibility for addressing current, ongoing harms associated with AI technologies right now. This strategy might be achieved by framing their public discourse in a way that shifts accountability and responsibility onto future scenarios, eluding immediate responses and actions. Lastly, it is noteworthy to mention that for DHA,

> *language is not powerful on its own; it is a means to gain and maintain power through the use powerful people make of it. This explains why the DHA critically analyses the language use of those in power who have the means and opportunities to improve conditions (Reisigl and Wodak, 2009: 88).*

In conclusion, Discourse-Historical Approach (DHA) is the appropriate research methodology for this study because its main goal lies in analysing discourses of powerful institutions and organizations in connection with heavily charged historical backgrounds, such as the technology industry and its historical and often unaccountable and unregulated impacts on society in the name of innovation.

Analytical Framework

Wodak's DHA provides a set of tools and principles for analysing texts. In this section, I will go over this structure in detail. According to Wodak and Reisigl (2009; 2015), DHA approaches textual meanings and structures through three dimensions, which are:

1. Thematic content or discourse topics: semantic macro-areas relevant to the discursive formation of discourses.

2. Discursive strategies: the strategies used in the discursive articulation of texts.

3. Linguistic means and context-dependent realizations: lexical units, argumentation schemes, and syntactical means for expressing unity, sameness, difference, singularity, continuity, change, autonomy, heteronomy, and so on (De Cillia, Reisigl, and Wodak, 1999: 160).

In order to correctly address these three layers of analysis, Wodak proposes five questions to further guide the investigation and, especially, to help identify discursive strategies as well as linguistic organizations. The five questions are (Wodak and Reisigl, 2009; 2015):

1. How are persons, objects, phenomena/events, processes and actions named and referred to linguistically?

2. What characteristics, qualities and features are attributed to social actors, objects, phenomena/events and processes?

3. What arguments are employed in the discourse in question?

4. From what perspective are these nominations, attributions and arguments expressed?

5. Are the respective utterances articulated overtly, intensified or mitigated?

For DHA, the concept of strategy is defined as 'a more or less intentional plan of practice [...] adopted to achieve a particular social, political, psychological, or linguistic goal' (Wodak and Reisigl, 2009; 2015). The level of strategic intentionality varies depending on the context and the genres of the texts. For example, a conversation between people about the dangers of AI might have a low level of intentionality, while a text on a company's website might show a much more conscious level of strategic intentionality, with the goal of framing the text in a specific way to convey a defined message (De Cillia, et al., 1999: 160).

Additionally, it is worth noting that the five guiding questions mentioned above make reference to specific macro-level discursive strategies proposed by Wodak, which are: nomination, predication, argumentation, perspectivization, and intensification/mitigation (Wodak and Reisigl, 2009; 2015).

However, there are more types of discursive strategies from which derive topics of argumentation and means of realization (Wodak, et al., 1999). Ultimately, what type of discourse strategies can be identified in a text will be linked to its context and topic.

To sum up, this framework's integration guarantees that the analysis is systematical, thorough, and sensitive to the language used strategically and contextually in OpenAI's discourse. Additionally, the three-dimensional approach outlined above is directly in line with the research question of the current study, which calls for a close and nuanced examination of OpenAI's discursive strategies, linguistic choices and sociohistorical context—all of which have the potential to influence regulatory and societal responses.

Sampling

The present research adopts a purposeful sampling strategy (Patton, 2015) and follows a single-case study method of inquiry. According to Creswell (1998) a case is defined as

> *an exploration of a 'bounded system' […] over time through detailed, in-depth data collection involving multiple sources of information rich in context. This bounded system is bounded by time and place, and it is the case being studied—a program, an event, an activity, or individuals.*

Although there is no one agreed definition of 'case' in the academic research field, I believe that Creswell characterization fits the focus of this study. Furthermore, a single-case study couples well with the chosen sampling strategy. According to Patton (2015),

> *qualitative inquiry typically focuses in depth on relatively small samples, even single cases (n = 1), selected for a quite specific purpose. […] The logic and power of purposeful sampling lies in selecting information-rich cases for in-depth study. Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the inquiry, thus the term purposeful sampling.*

To articulate the methods and procedures discussed within this single case study, it is necessary to provide readers with two types of rationales. The first rationale pertains to the selection of the case itself, while the second addresses the reasons and steps involved in defining the data collected for analysis.

Following the launch of its flagship product, ChatGPT, in late 2022, OpenAI has established itself as a prominent player in the AI field, gathering significant attention from media, governments, companies, academics, investors, and other stakeholders. As a result, OpenAI has gained substantial influence over narratives concerning AI, both present and future. A preliminary review of OpenAI's public communications revealed a recurring narrative emphasizing catastrophic and speculative AI risks, with comparatively few references to current ongoing harms. Hence, conducting a study regarding present and future AI discourses becomes particularly relevant, as AI technologies continue to evolve and pose present and long-term challenges for society.

In light of this context, the data collected for this study comes from three distinct sources and correspond to a time period ranging from May 2023 to June 2024.

- First, the 'Oversight of A.I.: Rules for Artificial Intelligence' hearing, hosted by the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law in May 2023, where OpenAI CEO Sam Altman testified. The entire hearing was transcribed, but only the relevant parts aligned with the research theme were considered for analysis. For instance, the testimonies of the other two witnesses were not taken into account.

- Second, two video interviews: an unedited interview with Sam Altman at the Bloomberg Technology Summit in June 2023 and an edited interview with Altman and OpenAI CTO Mira Murati, conducted by ABC News in March 2023. Both were obtained from YouTube using the search query 'OpenAI on risks' and transcribed for ease of analysis.

- Finally, the third dataset comprises the subpages listed under the secondary navigation menu in the 'Safety' category on the OpenAI website[5], captured for this analysis in June 2024. This dataset also includes a short video featured on one of the subpages, which was transcribed for the purpose of analysis.

The richness and diversity of the data collected align well with the principle of triangulation of data sources, a method that enhances the quality and credibility of the study (Patton, 2022; Wodak and Reisigl, 2015). This approach enables the detection of cohesiveness, or lack thereof, in OpenAI's narratives when addressing risks across various contexts. However, it is crucial to note that this

[5] https://openai.com/

research does not aim to systematically compare and contrast the discourse between sources. Rather, the objective of gathering an unusually large sample for CDA standards, was to capture a wide display of texts that would effectively represent OpenAI's stance on risks, avoiding skewed narrow perceptions.

In relation to the amount of material collected for each genre, a time or word count comparison between the sources is not useful in this case because of the big differences in format between the data types. So, in line with the purposeful sampling strategy, my goal was to capture a comprehensive and representative sample of each type of source based on the main topics posed by the research question.

The rationale for selecting these three specific genres is related to the audiences they are address to. By documenting diverse styles of language and expression it is possible to identify whether OpenAI addresses risk and safety differently when audiences change and how. For example, the hearing might exhibit a more cautious use of language, whereas the media interviews could reflect a more informal and speculative discourse. The website, on the other hand, is likely to demonstrate a highly intentional and strategic use of language. Additionally, the timeframe of the data set offers a temporal perspective that helps assess the consistency of OpenAI's public discourse on AI risks over a one-year span.

By this stage, I trust that I have clarified the ways in which influential organizations like OpenAI frame AI risks, which can significantly shape policy decisions, resource allocation, and regulatory measures. By applying the principles and tools outlined in this section, this case study examines whether prioritizing speculative future threats creates a disconnect in addressing present harms and assessing the effectiveness of existing policies—ultimately influencing society's broader relationship with AI technologies.

Ethics and Reflexivity

Critical Discourse Analysis (CDA) is often characterized as an engaged and committed form of intervention in social practices and relationships. However, this characteristic can apply to many other methodologies as well. According to Fairclough and Wodak (1997: 259), what distinguishes CDA from other qualitative and quantitative methods is that 'it intervenes on the side of the

dominated and oppressed groups and against dominating groups, and that it openly declares its emancipatory interest that motivates it.'

As a feminist and in line with D'Ignazio and Klein Data Feminism principles (2020), I view the declaration of the researcher's motivations and standpoint as both a necessary and powerful tool, especially in contrast to methodologies that emphasize objectivity. Also, as Gill (1996) observes, critical perspectives are essential for 'challenging authoritarian authorities and drawing attention to the status of their own texts as constructions.'

It is important to note that criticality does not imply a lack of quality, effectiveness, or operational tools for analysis. On the contrary, exploring the complexities of discourse as a socio-cultural practice requires upholding high research standards and employing a comprehensive research design and methodological framework like Wodak's DHA. However, regardless of the analytical tools used, the ethics and reflexivity of the researcher remain crucial in any study. This includes recognizing one's own standpoint, engaging with the data with constant skepticism, and clearly understanding the limitations of the methodology. Also, it is important to take into account that the process of data collection, even though it was guided by a methodological justification and rationale, it is the result of what algorithms showed me based on my interests and past online history.

Critical Discourse Analysis (CDA), and by extension Discourse-Historical Analysis (DHA), have their limitations. For instance, the flexibility and holistic nature of CDA approaches can lead to varying conclusions from the same data, affecting the reproducibility of the study and limiting the ability to make broader claims. Therefore, researchers must acknowledge from the outset that 'different analytical approaches will yield different kinds of findings based on distinct analysis procedures and priorities' (Patton, 2015).

Additionally, concerns have been raised about the representativeness of findings in CDA studies. While it might be tempting to generalize conclusions to a broader population, this would miss the essence of the method. The primary goal of CDA and DHA is to stimulate discussion and bring overlooked viewpoints and themes to the public agenda—issues that are often considered non-problems and, if left unchallenged, may remain unexplored and rendered in societal status quos.

ANALYSIS AND DISCUSSION

I begin this section by outlining the three primary topics that emerged from the analysis of the sampled texts concerning OpenAI's discourse on AI risk

1. The discourse construction of AI and OpenAI through contrasting comparisons
2. The narrative of society's collective power to convey trust and responsibility
3. The narrative of a AI as the only solution to society's most challenging problems

The following subsections present the findings of a thorough analysis of how OpenAI's communicate its stance on AI risks and safety. The findings reflect the interpretations and conclusions derived by applying Wodak's DHA tools and principles to unpack the strategies used to construct these narratives and their implications in conjunction with the theoretical and conceptual framework to answer the proposed research question.

The Dichotomy of Potential Danger: Constructing AI as a Double-Edged Sword

OpenAI's public discourse around AI risks, as illustrated by the following quotes, reflects a strategic construction that oscillates between presenting AI as a source of tremendous benefits and as a potential cause of catastrophic risks. This dual narrative shapes public perception and policy-making in specific ways, often by juxtaposing utopian and dystopian possibilities.

A significant aspect of this discourse is the construction of AI within the dichotomy of potential versus threat. OpenAI frequently emphasizes this duality, pairing AI's potential to solve humanity's most significant challenges with the risks of catastrophic consequences. For instance, the following statements from its website capture this dichotomy:

> *'Superintelligence will be the most impactful technology humanity has ever invented, and could help us solve many of the world's most important problems. But the vast power of superintelligence could also be very dangerous, and could lead to the disempowerment of humanity or even human extinction' (Appendix A 4a).*

> *'Very quickly, we can end up in a place where machines are far more capable than science and they can help us solve very hard scientific problems that humans are not capable of solving themselves. So, what's really interesting is as the AI systems get more capable, they*

> *don't automatically become better at doing what humans want. In fact, sometimes they become less inclined to following human intentions' (Appendix A 4c).*

This construction serves several discursive purposes. First, it legitimizes the pursuit of advanced AI, positioning it as essential to future progress. This makes it difficult for critics to argue against AI's development without appearing to oppose technological advancement and innovation, which are core national values in the United States. Second, it implicitly positions OpenAI as a responsible and trustworthy entity by presenting the dangers as existential threats, counterbalanced by the potential for tremendous benefits. Third, this dual narrative evokes popular sci-fi sociotechnical imaginaries of ubiquitous and powerful AI, while completely omitting the current ongoing harms. This implies that the risks of AI systems are only a threat in the future.

Continuing with the analysis of the sample texts, a pattern of obscuring current harms by emphasizing future existential risks becomes evident in OpenAI's communications. The quotes provided primarily focus on hypothetical scenarios where AI could cause significant harm, leading to disastrous outcomes. For example, the comment made by Altman during the Congressional Hearing,

> *If this technology goes wrong, it can go quite wrong...we try to be very clear-eyed about what the downside cases and the work that we have to do to mitigate that… (Appendix A 1),*

shifts the focus toward extreme potential future risks. His emphasis on speculative dangers correlates directly with downplaying the immediate and ongoing harms of AI, such as the exacerbation of bias, disinformation, economic disruption, and job exploitation, which are only briefly mentioned and discussed in the text samples.

Additionally, OpenAI employs vague language and uncertainty as a discursive strategy when addressing existential threats and AI benefits, while simultaneously obscuring ongoing negative consequences and overhyping future positive outcomes. The strategy of using vague language and emphasizing uncertainty creates ambiguity, allowing OpenAI to avoid making definitive statements or commitments about AI's effects and implications. This could lead to a societal understanding that underestimates the importance and urgency of addressing ongoing harms, focusing instead on speculative future scenarios that are less tangible and harder to address. Consequently, this focus on

future risks could lead to policies that prioritize long-term existential risk mitigation over the regulation of present-day AI applications, inadequately addressing significant social harms.

Another recurring theme in OpenAI's discourse is the 'alignment problem', where the challenge of ensuring AI systems adhere to human values and intentions is framed as critical for the preservation of society. This quote from an OpenAI website video:

> *I think solving this problem is of critical importance if we want life on Earth to go well. Like humans, when machines learn, they make mistakes. And so, the question is, how do we prevent machines from making mistakes that have significant consequences? Even seemingly obvious values like telling the truth, the system actually has to be incentivized to do and has to want to tell you the truth. Even today, we can't peer into the depths of the neural net and understand what's happening inside the mind of the machine. So how do we make sure that the system actually acts in accordance with human intentions and in accordance with human values? (Appendix A 4c),*

constructs a narrative where AI's increasing capability is directly linked to the potential end of humanity. This is achieved by reinforcing sociotechnical imaginaries of AI as an entity with its own agency and will—an entity that '*has to want to tell you the truth'*—and by applying the discursive strategy of anthropomorphizing AI systems, directly connecting them to human behaviors. This suggests that AI needs to be '*incentivized*' or that '*like humans, AI makes mistakes.*'

Through these quotes, it becomes evident how OpenAI's discourse strategically constructs AI's dual potential—its immense benefits and catastrophic risks—to influence societal perceptions and policy directions. OpenAI employs strategic ambiguity, where the future is presented as both promising and catastrophic. This serves to maintain public support for AI development while crafting an image of responsibility and trustworthiness, framing itself as a cautious, ethical actor in AI development. The power of these narratives lies in their ability to shape societal understanding of AI and policy environment on speculative risks and benefits rather than prioritizing immediate, actionable issues, potentially slowing efforts to address the current challenges posed by AI technologies.

Collective Power: Shared Responsibility and Diffused Accountability

One of the central strategies in OpenAI's discourse is the invocation of society´s collective power and the importance of its role in deciding and shaping AI technologies. This is suggested implicit and

explicitly in diverse ways. For instance, the following quote made by a Senator to Sam Altman on the Congressional Hearing: *'I think what's happening today in this hearing room is historic, […] companies telling the government 'Stop me before I innovate again' message...' (Appendix A 1)*, reflects the perception of the US government of OpenAI as a rare case, implying that private companies usually avoid being regulated. In context, this is significant, because it makes a clear cut between the perception that the Senators have of other Big Tech Companies (i.e. Facebook, Google, etc), that it are characterized as being too powerful, irresponsible and unwilling to take responsibility for its actions, as well as the US government being too slow and inexpert to recognize the repercussions of an unregulated tech industry on society, notions that are address on this Hearing (Appendix A 1).

This rhetorical move positions OpenAI as a responsible actor seeking to collaborate with the government, thus spreading the responsibility for AI's risks and benefits across both the private and public sectors. By framing the need for regulation as a shared endeavor, OpenAI implies that the obligation of managing AI's potential dangers is a collective one, not just the responsibility of the creators.

The narrative of shared accountability is further reinforced by the statement, *'Certainly, companies like ours bear a lot of responsibility...but tool users do as well' (Appendix A 1).* This quote explicitly distributes responsibility to both AI developers and users, suggesting that the impacts of AI are a product of how it is used, not just how it is created. This construction serves to partially offload the accountability for AI risks onto users, who are depicted as active agents with a significant role in ensuring that AI tools are used responsibly. By emphasizing user responsibility, OpenAI subtly deflects some of the scrutiny away from the potential harms embedded within the AI systems themselves, instead focusing on the behaviors and decisions of those who interact with these systems.

OpenAI's discourse also frequently invokes the idea of societal involvement in setting the ethical and operational boundaries of AI. The statement: *'What these systems get aligned to, whose values, what those bounds are...that is somehow set by society as a whole' (Appendix A 1),* underscores the notion that the ethical alignment of AI should be a democratic process, involving broad societal input. This narrative not only serves a democratizing factor for the control over AI but also implies that the consequences of AI, whether positive or negative, are the result of collective societal decisions. By promoting the idea that society as a whole should determine AI's alignment values and boundaries, OpenAI

effectively spreads the responsibility for AI's outcomes across governments and societies, thereby diluting the company's sole accountability, even on a global scale.

In line with the above, the appeal to democratic values is closely linked to the need of US leadership in AI, as seen in following statement: *'It is essential that powerful AI is developed with democratic values in mind, and this means that US leadership is critical' (Appendix A 1).* On one hand, it aligns the development of AI with widely accepted democratic principles, thereby fostering public trust. On the other hand, it positions the United States, and by extension OpenAI, as the rightful leaders in the global AI landscape, subtly suggesting that other nations will have to follow its example, especially less powerful ones. This narrative helps to consolidate US power while simultaneously promoting a narrative of shared responsibility, as the involvement of a democratic society implies collective participation of private and public spheres for AI's beneficial development and deployment.

In addition to framing AI development as a collective responsibility, OpenAI's discourse also emphasizes the importance of public input and customization, as seen in: *'we are working on gathering public input...within these hard bounds, you can have a lot of choice in having your own AI represent your own beliefs and your own values' (Appendix A 2).* This rhetoric suggests that not only is AI development a collective effort, but individual users have significant control over how AI reflects their personal values. This narrative serves to both empower users and distribute the accountability for AI's actions more broadly, implying that any misalignment or harm caused by AI could partly be attributed to the choices made by individual users.

This analysis reveals that OpenAI's discourse strategically constructs a narrative that balances the need for control through the promotion of a collective responsibility discourse. This approach helps to obscure accountability of current AI harms by making it seem the users are greatly involved in how AI systems are used and deployed, and also to shift the focus onto future scenarios and the collective efforts needed to mitigate them. By framing AI development as a democratic, collective process that requires the input and responsibility of all stakeholders —developers, users, governments, and society in general — OpenAI effectively diffuses responsibility and accountability. This diffusion can lead to a societal understanding of AI where bad actors are the ones to blame for current AI harms. Also, the implications for policy-making are significant, as this narrative may result in regulatory frameworks that prioritize future risk mitigation over actively review legal frameworks

that can address present-day harms, thereby allowing current issues related to AI deployment to persist under the guise of a shared, yet somewhat diluted, responsibility.

Framing the Future: Positioning Technologies as Inevitable and Necessary

By critically analysing OpenAI discourse it becomes evident that the tech company systematically positions AI as an inevitable and essential development in human history. This is evident in several texts. For instance, the statement:

> *Like there are a lot of people that talk about AI as like the last technological revolution, I suspect that you know, from the other side, it'll look like the first, like the other stuff will be so small in comparison. I think the whole thing of like technological revolutions is sort of dumb, because my understanding has always been it's just one long, continuous one, but it is this continuing, exponential. And so what, what will be enabled in the stuff we can't even imagine on the other side, we will have way too much to do if you want, if you want to just sit around and do nothing. That would be fine, too. (Appendix A 3)*

frames AI as a continuation of an inevitable, ongoing technological revolution. This portrayal minimizes the possibility of rejecting or significantly altering the path of AI development by presenting it as part of an unstoppable historical process.

Similarly, through the the assertion:

> *Now GPT-4 will, I think, entirely automate away some jobs, and it will create new ones that we believe will be much better. This happens again. My understanding of the history of technology is one long technological revolution, not a bunch of different ones put together. But this has been continually happening. We, as our quality of life raises, and as machines and tools that we create can help us live better lives, the bar raises for what we do and our human ability and what we spend our time going after goes after more ambitious, more satisfying projects. (Appendix 1 A),*

Open AI suggests that AI is simply the latest step in a continuous process of human progress and uses the discursive strategy of 'history as a teacher' reminding us that we have survived other technological innovations and have achieved a 'better life' despite the harms and risk posed. This framing of technology and AI reinforces the idea that resistance or other alternative solutions are useless because technology has proven to be the right path in the past, as well as ranking technological advancement as the most natural and legitimate solution to any problem.

Moreover, the emphasis on superintelligence in the following quote:

> *Here we focus on superintelligence rather than AGI to stress a much higher capability level. We have a lot of uncertainty over the speed of development of the technology over the next few years, so we choose to aim for the more difficult target to align a much more capable system. (Appendix A 4a)*

serves to show how OpenAI manages its AI approach as future-oriented, investing talent and resources in a more advanced stage of AI development. By setting the narrative focus on superintelligence, OpenAI frames the present challenges and risks as necessary trade-off towards a much greater, and more necessary, future goal. This not only shifts attention away from current concerns but also implies that current sacrifices are justified by the potential of future advancements.

The way that OpenAI's language often explicitly shifts the focus away from present-day harms by emphasizing the management of future risks can be seen in the following statement:

> *We are investing in the design and execution of rigorous capability evaluations and forecasting to better detect emerging risks. In particular, we want to move the discussions of risks beyond hypothetical scenarios to concrete measurements and data-driven predictions. We also want to look beyond what's happening today to anticipate what's ahead. This is so critical to our mission that we are bringing our top technical talent to this work (Appendix A 4b.).*

This future-oriented discourse downplays current harms by implying that the real dangers lie ahead, thus justifying the present trade-offs as necessary for long-term safety. Additionally, the claim *'We believe that the benefits of the tools we have deployed so far vastly outweigh the risks...' (Appendix A 1),* emphasizes the vast current benefits of AI, positioning them as already outweighing current risks. This rhetoric minimizes the significance of present-day harms, suggesting that these are trivial in comparison to the promised future benefits, thus implying a cost-benefit analysis where the potential positive outcomes of AI are deemed worth the current and near-term sacrifices.

The narrative that AI is an inevitable force and that its future benefits justify present risks could lead to a societal consensus that current harms are tolerable. This perception can weaken efforts to critically evaluate and address the immediate negative impacts of AI, such as job displacement, job precariousness, privacy and copyright violations, and bias in AI systems. As OpenAI is considered by policymakers to be a responsible actor in the tech industry, these narratives and discursive

strategies might cause to prioritize long-term risk management regulations over immediate regulatory interventions in ingoing harms due to the framing of near future existential risks scenarios as the most significant concern. This could result in a regulatory environment that is less responsive to the current social, economic, and ethical challenges posed by AI technologies.

Moreover, the emphasis on a continuous technological revolution marginalizes voices that have historically advocated for a more cautious or alternative approach to AI development. Currently OpenAI discourse presents a narrow view of progress, sidelining those who demand for ethical, social, or economic considerations to take precedence over rapid technological advancement and innovation.

CONCLUSION

In conclusion, this dissertation reveals that OpenAI's public discourse on AI risks and safety strategically emphasizes future existential threats while downplaying current real harms. By framing AI as a dual-edged sword with immense potential benefits and catastrophic risks, OpenAI positions itself as a responsible and ethical leader, thus influencing societal understanding and policy-making in ways that prioritize long-term speculative dangers over immediate, actionable issues. The narratives of shared responsibility and the inevitability of technological progress further diffuse accountability, leading to a societal and regulatory focus on future risks rather than addressing the pressing harms of AI today.

The findings discussed in the previous section suggest that OpenAI's discourse may contribute to a regulatory environment that inadequately addresses ongoing social, economic, and ethical challenges posed by AI. As a result, there is a risk that policies will be shaped more by fears of hypothetical future scenarios than by the need to mitigate current harms, such as bias, disinformation, and job displacement.

Further research could explore how these discursive strategies impact public opinion and policy development across different global contexts, particularly in regions with varying levels of technological development and regulatory frameworks. Additionally, examining how other key AI players construct similar narratives could provide a broader understanding of the tech industry's role in shaping societal and policy responses to AI. Finally, longitudinal studies could investigate how

these discourses evolve over time and their long-term implications for both AI governance and public trust in AI technologies.

REFERENCES

Ballatore, A. and Simone, N. (2023) Technological failures, controversies and the myth of AI, pp. 237-244 in S. Lindgren (ed.) *Handbook of Critical Studies of Artificial Intelligence*, Cheltenham: Edward Elgar Publishing.

Bender, E. M. (2023) Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence, written testimony of Dr. Emily M. Bender before the U.S. House Committee on Science, Space, and Technology, U.S. House of Representatives, URL: https://democrats-science.house.gov/imo/media/doc/Dr.%20Bender%20-%20Testimony.pdf.

Bender, E. M., Gebru, T., McMillan-Major A., and Shmitchell S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21): 610–623, New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bones, H., Ford S., Hendery R., Richards K., and Swist, T. (2021) In the Frame: The Language of AI, *Philosophy & Technology* 34 (1): 23–44. https://doi.org/10.1007/s13347-020-00422-7.

Brevini, B. (2021) Creating the Technological Saviour: Discourses on AI in Europe and the Legitimation of Super Capitalism, pp. 145-160 in P. Verdegem (ed.) *AI for Everyone? Critical Perspectives*. University of Westminster Press. https://www.jstor.org.gate3.library.lse.ac.uk/stable/j.ctv26qjjhj.11.

Barbrook, R., and Cameron A. (1996) The Californian Ideology, *Science as Culture* 6 (1): 44–72. https://doi.org/10.1080/09505439609526455.

Couldry, N. and Mejias, U. A. (2019) Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject, *Television & New Media* 20 (4): 336–349. https://doi.org/10.1177/1527476418796632

Creswell, J. W. (1998) *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Creech, B. and Maddox, J. (2024) Thus Spoke Zuckerberg: Journalistic Discourse, Executive Personae, and the Personalization of Tech Industry Power, *New Media & Society* 26 (7): 4201–18. https://doi.org/10.1177/14614448221116344.

De Cillia, R., Reisigl, M. and Wodak, R. (1999) The Discursive Construction of National Identities, *Discourse & Society* 10(2): 149-173. https://doi-org.gate3.library.lse.ac.uk/10.1177/0957926599010002002.

D'Ignazio, C. and Klein, L. F. (2020) *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Fairclough, N. (1993) Critical Discourse Analysis and the Marketization of Public Discourse: The Universities, *Discourse & Society* 4 (2): 133–68. https://doi.org/10.1177/0957926593004002002.

Fairclough, N. and Wodak, R. (1997) Critical Discourse Analysis, pp. 258-284 in T. A. Van Dijck (ed.) *Discourse as Social Interaction*, SAGE Publications Ltd.

Ferri, G. and Gloerich, I. (2023) Risk and Harm: Unpacking Ideologies in the AI Discourse, *Proceedings of the ACM Conference on Conversational User Interfaces (CUI),* Eindhoven, Netherlands, 1-6. New York, NY: ACM. https://doi.org/10.1145/3571884.3603751.

Gill, R. (1996) Discourse Analysis: Practical Implementation, in J. T. E. Richardson (ed.) *Handbook of Qualitative Research Methods for Psychology and the Social Sciences*, Leicester: British Psychological Society.

Gill, R. (2000) Discourse Analysis, pp. 173-190 in M. W. Bauer and G. Gaskell (eds) *Qualitative Researching with Text, Image and Sound*. SAGE Publications Ltd. https://doi.org/10.4135/9781849209731.

Haupt, J. (2021) Facebook Futures: Mark Zuckerberg's Discursive Construction of a Better World, *New Media & Society* 23 (2): 237–257. https://doi.org/10.1177/1461444820929315.

Hepp, A., Schmitz A. and Schneider, N. (2023) Afterlives of the Californian Ideology: Tech Movements, Pioneer Communities, and Imaginaries of Digital Futures, *International Journal of Communication* 17: 4142–4160.

Hicks, M. T., Humphries J., and Slater J. (2024) ChatGPT is Bullshit, *Ethics and Information Technology* 26(2): 1-10. https://doi.org/10.1007/s10676-024-09775-5.

Jasanoff, S. and Kim, S. (eds) (2015) *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago, IL: University of Chicago Press. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=2130453.

Jasanoff, S. (2015) Future Imperfect: Science, Technology, and the Imaginations of Modernity, pp. 1-33 in Jasanoff, S. and Kim, S. (eds) *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*, Chicago IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226276663.003.0001.

Mager, A. and Katzenbach, C. (2021) Future Imaginaries in the Making and Governing of Digital Technology: Multiple, Contested, Commodified, *New Media & Society* 23(2): 223–236. https://doi.org/10.1177/1461444820929321.

Morozov, E. (2013) *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems That Don't Exist*. New York, NY: Allen Lan

Mosco, V. (2004) *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge, MA: MIT Press.

Patton, M. Q. (2015) *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (Fourth edition.). SAGE.

Rehak, R. (2021) The Language Labyrinth: Constructive Critique on the Terminology Used in the AI Discourse, pp. 87-102 in P. Verdegem (ed.) *AI for Everyone? Critical Perspectives.* University of Westminster Press. https://www.jstor.org.gate3.library.lse.ac.uk/stable/j.ctv26qjjhj.8.

Reisigl, M. and Wodak, R. (2009) The Discourse-Historical Approach (DHA), pp. 87-121 in R. Wodak and M. Meyer (eds) *Methods for Critical Discourse Analysis*, London: SAGE (2nd revised edition).

Richter, V., Katzenbach, C. and Schäfer, M. S. (2023) Imaginaries of Artificial Intelligence, pp. 209-222 in S. Lindgren (ed.) *Handbook of Critical Studies of Artificial Intelligence*, Cheltenham: Edward Elgar Publishing.

U.S. Department of Justice (2020) Department of Justice's Review of Section 230 of the Communications Decency Act of 1996, *U.S. Department of Justice*, 17 June, URL: https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996. [Last consulted August, 2024].

Verdicchio, M. (2023) Marking the Lines of Artificial Intelligence, pp. 245-253 in S. Lindgren (ed.) *Handbook of Critical Studies of Artificial Intelligence*, Cheltenham: Edward Elgar Publishing

ten Oever, N. (2021) This Is Not How We Imagined It: Technological Affordances, Economic Drivers, and the Internet Architecture Imaginary, *New Media & Society* 23(2): 344–362. https://doi.org/10.1177/1461444820929320.

Wodak, R. and Reisigl, M. (2015) The Discourse-Historical Approach (DHA), pp. 23-61 in R. Wodak and M. Meyer (eds) *Methods of Critical Discourse Studies*, London: SAGE Publications Ltd, 3rd edition.

Wodak, R., de Cillia, R., Reisigl, M. And Liebhart, K. (1999) *The Discursive Construction of National Identity*. Edinburgh: Edinburgh University Press.

Wyatt, S. (2021) Metaphors in Critical Internet and Digital Media Studies, *New Media & Society* 23(2): 406–416. https://doi.org/10.1177/1461444820929324.

APPENDICES

Appendix A – Sample Texts

1. Subcommittee Hearing - Oversight of A.I.: Rules for Artificial Intelligence:

U.S Senate Committee on the Judiciary. 16 May, 2023. Subcommittee on Privacy, Technology, and the Law. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence.

2. ABC News interview to Sam Altman (CEO) and Mira Murati (CTO) of OpenAI:

ABC News. 23 March, 2023. *OpenAI CEO, CTO on Risks and How AI Will Reshape Society*. https://www.youtube.com/watch?v=540vzMlf-54.

3. Bloomberg Technology Summit interview to Sam Altman:

Bloomberg Live. 22 June, 2023. *OpenAI CEO Sam Altman on the Future of AI*. https://www.youtube.com/watch?v=A5uMNMAWi3E.

4. OpenAI website:

a. OpenAI. n.d. Introducing Superalignment, Safety. URL: https://openai.com/index/introducing-superalignment/. [Last consulted 22 June, 2024].

b. ———. n.d. Preparedness, Safety. URL: https://openai.com/preparedness/. [Last consulted 22 June, 2024].

c. ———. n.d. Safety & Responsibility, Safety. URL: https://openai.com/safety/. [Last consulted 22 June, 2024].

d. ———. n.d. Safety Standards, Safety. URL: https://openai.com/safety-standards/. [Last consulted 22 June, 2024].

e. ———. n.d. Safety Systems, Safety. URL: https://openai.com/safety-systems/. [Last consulted 22 June, 2024].

Appendix B: Coding Table and Annotations Example

Safety Overview Intro Video | OpenAI Audio Transcript

Safety Overview Intro Video | OpenAI Audio Transcript

Retrived from: https://openai.com/safety/ on 15 August 2024

00:12
Today, we can use AI to write poetry, compose music, write computer programs.
Very quickly, we can end up in a place where machines are far more capable than say science and they can help us solve very hard scientific problems that humans are not capable of solving themselves.

00:36
So what's really interesting is as the AI systems get more capable, they don't automatically become better at doing what humans want. In fact, sometimes they become less inclined to following human intentions. This is what we call the alignment problem.

00:52
I think solving this problem is of critical importance if we want life on Earth to go well.
Like humans, when machines learn, they make mistakes. And so the question is, how do we prevent machines from making mistakes that have significant consequences.

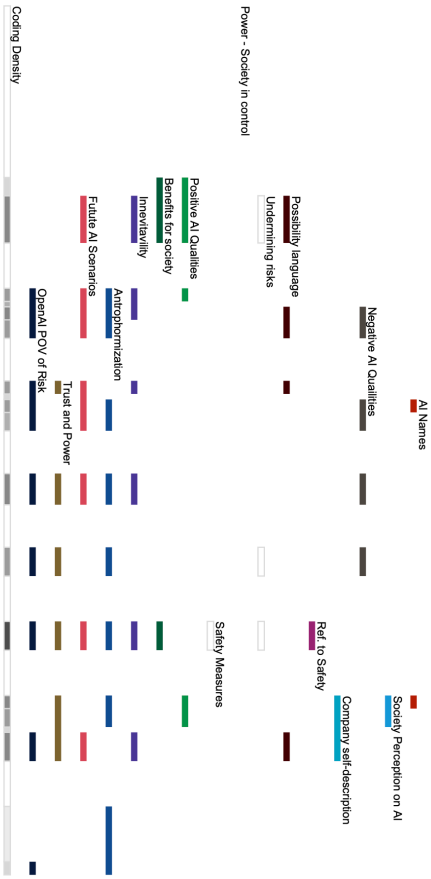
01:13
Even seemingly obvious values like telling the truth, the system actually has to be incentivized to do and has to want to tell you the truth.

01:22
Even today, we can't peer into the depths of the neural net and understand what's happening inside the mind of the machine.

01:30
So how do we make sure that the system actually acts in accordance with human intentions and in accordance with human values?

01:39
For the first time in the history of AI, we have this very powerful, large language models like GPT-3 that has such linguistic competence that sometimes it's indistinguishable from what humans can produce.
But technically, it's not a trivial problem to figure out how to get these machines to do the things that we want them to do.

02:03
So for example, if you ask GPT-3, please explain the moon landing to a five year old, it will try to guess what the pattern is, and might say something like, how do you explain the concept of infinity to a five year old?
Explain humor, comedy parenting to a five year old and so on.



Where do babies come from? What is war? And so it's like trying to guess the pattern of what we're getting at, but that's not actually what you wanted, right? You wanted an actual explanation. And so we have to align GPT-3 to follow instructions. And we do that by designing systems that learn from human feedback.

02:43
As a first step, we show the model what it means to follow instructions, and so we have our researchers provide a bunch of demonstrations of questions and answers, and then as a second step, we have a human look at a bunch of responses and say, I like this one better than that one, and so on. And little by little, the system learns to follow instructions as humans want.

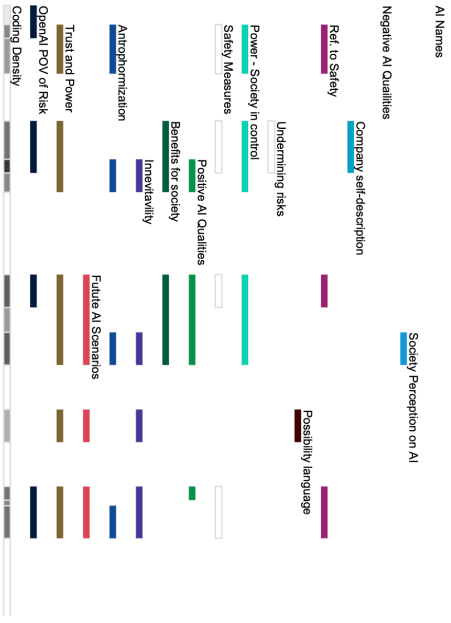
"In July 20 1969 two astronauts did something no one have ever done before" (background audio of adult and child interacting).

So using human feedback, we can align the system to follow instructions, and that makes it more useful, more reliable and more trustworthy.

Then you end up with a collaboration between humans and AI. We teach AI our individual values, and AI helps us, in turn, by living better more fulfilling lifes.

03:41
AI is going to play a larger and larger role in our lives, and that raises the question, where are we going with all this, and what's going to happen in the future?

03:50
As these systems become more powerful, alignment will become even more critical.
It's probable that AI systems will become a part of everyday life, and the key is to ensure that these machines are aligned with human intentions, and human values.



Codes - OpenAI Risk Discourse

Name	Description	Sources	References
AI Names	Ways to name AI (referencing to AI directly)	5	21
Antrophormization		6	25
Benefits for society	What AI can helo humans do	7	41
Company self-description		5	21
Fututе AI Scenarios		8	54
Innevitavility		8	56
Media Public Perception		2	22
Negative AI Quailities	According to OpenAI	6	17
OpenAI POV of Risk		8	79
Safety Measures		8	43
Undermining risks	By OpenAI views	7	26
Positive AI Qualities	Words and phrases that describe positively AI technologies and chat gpt	6	32
Positive Action Verb	Related to AI - What AI does	1	6
Possibility language		6	42
Ref. to Safety	As Ideology or a goal	5	19
Society Perception on AI		4	19
Trust and Power		8	93
Power - Society in control	Linked to accountability and responsibility	7	33