Moving Forward in AI Development: A Constructivist Grounded Theory Approach with Systems Thinking Lens

Meng Ma, Haskayne School of Business, University of Calgary
Giovani J.C. da Silveira, Haskayne School of Business, University of Calgary

Abstract

The fast evolution of artificial intelligence has introduced significant opportunities and challenges, necessitating a comprehensive understanding of its complexities. This study employs Constructivist Grounded Theory to explore expert perspectives on artificial intelligence (AI) development, complemented by the DSRP framework - Distinctions, Systems, Relationships, and Perspectives - to analyze emerging themes. Starting with preliminary, sensitizing concepts based on theoretical background, data collection and analysis was used to develop and refine definitive themes iteratively, with themes guiding the search for relevant sources until saturation. This analysis revealed four key themes: AI capability, the impact of AI, AI alignment, and AI agency. From these themes, two theories emerged: the AI alignment theory, emphasizing transparency, decentralization, and a shift towards human-AI progress; and the symbiotic relationship theory, highlighting the asymmetry of intelligence between humans and AI and the delegation of agency. These theories provide meaningful insights for understanding AI development dynamics, informing both academic research and practical policymaking.

1. INTRODUCTION

The field of artificial intelligence (AI) development has rapidly evolved, presenting both unprecedented opportunities and significant challenges. As AI systems become more integrated into various aspects of human life, from healthcare to finance to daily decision-making, understanding the complexities of AI development is essential (Russell et al., 2015). These complexities include not only technical advancements but also the ethical, social, and economic implications (Hacker, Engel, & Mauer, 2023). To navigate this intricate landscape, it is crucial to engage with diverse perspectives and develop frameworks that can comprehensively address the multifaceted nature of AI.

In this paper, we employ Constructivist Grounded Theory (CGT) to explore the perspectives of experts in the field of AI. We started with a few sensitizing concepts (Blumer 1954) based on literature background. Following the inductive process of grounded theory, our data collection was iterative and evolving. Sensitizing concepts helped with the identification of initial sources, and preliminary themes. These themes guided the search for additional sources, which were filtered to ensure they were not redundant or irrelevant. Coding new data often resulted in the emergence of new themes,

or refining of preliminary themes, prompting further data collection. This cycle continued until existing themes appeared to be saturated, and new themes began to deviate significantly from the core relevance to AI. This grounded approach allows for a comprehensive and dynamic exploration of the field, rooted in real-world experiences and insights (Charmaz, 2006).

To further analyze the themes identified through CGT, we applied the DSRP framework - Distinctions, Systems, Relationships, and Perspectives - initially proposed by Cabrera, Colosi, and Lobdell (2008) as a formalism of systems thinking. The DSRP framework brought significant benefits to our analysis by capturing complexity and providing a structured lens for examining components and dynamics within the data (Cabrera and Colosi, 2008). For instance, making distinctions (D) helps define problems clearly, while keeping fluid boundaries reminds us that the meanings of terms can change depending on different users. Similarly, when discussing concepts at different levels of abstraction using the system-component pair (S), the flexibility to move between these levels can reveal emergent features that might be missed otherwise.

The framework also unveils relationships (R) by explicitly focusing on how different stakeholders (e.g., developers, policymakers) interact and potentially clash regarding AI development, exploring the social dynamics shaping the field. Considering diverse perspectives (P) allows us to analyze how various viewpoints (ethical, technical, economic) influence the discourse on AI development, which is crucial in a field marked by ongoing debates and value considerations. The flexibility and adaptability of DSRP allow themes and connections to emerge organically during the coding process, enriching theory building by accounting for the interplay of various factors and stakeholder influences.

Moreover, DSRP bridges the gap between abstraction and concreteness by enabling movement between conceptual distinctions and practical systems-level analysis (Cabrera and Colosi, 2008). This is particularly valuable in AI development, where abstract concepts like “agency” have real-world implications for system design and decision-making. Finally, the focus on systems encourages the identification of unforeseen connections, leading to new insights into how seemingly separate aspects of AI development are interconnected.

By integrating CGT with the DSRP framework, this study uncovers four key themes: AI capability, AI impact, AI alignment, and AI agency. From these themes, two theories emerge: the AI alignment theory and the symbiotic relationship theory. The AI alignment theory highlights the importance of transparency, decentralization, and a shift towards common human-AI progress while critiquing top-down government regulations. The symbiotic relationship theory addresses the asymmetry of intelligent capability between humans and AI and the delegation of agency towards AI, emphasizing the potential for a symbiotic human-AI relationship. These theories are significant in both research and practice, offering a deeper understanding of AI development dynamics and guiding policymaking and strategies moving forward.

2. BACKGROUND

2.1. Artificial Intelligence

AI involves creating machines, particularly computers, that emulate human intelligence. AI has advanced significantly from its beginnings with knowledge-based systems to the present era dominated by deep learning and transformer-based generative models (Roser, 2024). These advancements have dramatically improved AI's capabilities, enabling machines to process information much more efficiently than humans, and to perform interactive tasks that closely mimic human input and output (Familoni & Onyebuchi, 2024).

Initially, AI focused on algorithms that simulated human thinking processes, with the aim of programming machines to reason like humans. Early AI, constrained by the limited hardware of the time, was dominated by rule-based or expert systems (Hayes-Roth & Jacobstein 1994). These systems relied on hard-coded rules provided by experts to simulate human expertise in specific areas. However, these systems struggled with scalability and adaptability, leading researchers to seek more flexible and powerful approaches (Guo, Pan, & Heflin 2004).

This search led to the development of statistical methods and machine learning, where algorithms learned patterns from data. Despite their promise, these methods had limitations, such as the need for extensive feature engineering and challenges in handling non-linear data (Malik, 2020). The breakthrough came with deep learning, which utilized neural networks inspired by the human brain to extract features and learn complex patterns. The implementation of deep learning with architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) revolutionized fields such as image and video recognition, natural language processing, and time series analysis (Dhruv & Naskar, 2020; Khan, 2020). More recently, transformer-based models have further advanced AI, significantly improving tasks like language translation and text generation by capturing long-range dependencies in data (Vaswani et al., 2017). These models' versatility has led to their application in diverse areas, from content creation to scientific research, marking the latest leap in AI's evolution.

2.2. Sensitizing Concepts

Discussions and concerns about recent AI breakthroughs have centered on several key topics. Ethics, regulation, and AI safety have been major focuses, with debates about bias, fairness, privacy, and the need for comprehensive governance frameworks to ensure ethical AI development and deployment (Pereira et al., 2023). The impact of AI on society has also been a prominent issue, with fears of job displacement and the need for reskilling and upskilling the workforce to adapt to new roles created by AI technologies (Sofia et al., 2023). Additionally, advancements in AI technologies and their applications in various fields have sparked significant interest and discussion (Richardson & Heck, 2023).

Many of these topics originate from a deeper understanding of AI's fundamental features and philosophies (Youheng, 2023). Core concepts such as AI alignment and the exploration of AI's capabilities and limitations underpin these discussions. These foundational ideas serve as lower-level abstractions that inform emergent topics within the AI realm (Müller, 2024). For instance, understanding AI alignment involves philosophical questions about ensuring AI systems act in accordance with human values, which is crucial for addressing concerns about AI safety and ethics. Similarly, the exploration of AI's capabilities helps identify the potential and limits of current AI projects, shaping discussions about their practical applications and implications.

These fundamental topics of understanding became the "*sensitizing concepts*" (Blumer, 1954) in our grounded theory approach, guiding the initial phases of theme coding and data collection. Sensitizing concepts are provisional rather than definitive constructs (Blumer, 1954). They provide an anchor of reference and general direction, allowing the researcher to remain open to emerging themes while grounded in a conceptual base. By focusing on core ideas such as alignment, impact, and capability exploration, we set the stage for a comprehensive analysis of AI-related discussions, making sure that our research captures the depth and breadth of the ongoing discourse.

Thus, data saturation in our study was achieved by mapping new data sources initially to sensitizing concepts and soon to preliminary themes that followed. Potential new data were categorized based on their relationships with existing themes. If the new data were not relevant enough to fit within the existing categories, they were discarded. This approach ensured that our study remained focused on the essential aspects of AI discussions, capturing the critical themes while filtering out less pertinent information.

3. METHODOLOGY

3.1. Constructivist Grounded Theory

Grounded Theory (GT) is a qualitative research methodology that aims to develop theories grounded in systematically gathered and analyzed data. It was originally developed by sociologists Barney Glaser and Anselm Strauss (1968). The primary objective of GT is to allow theories to emerge from the data itself rather than testing existing theories or hypotheses. This approach emphasizes an inductive process where the researcher collects and analyzes data simultaneously, constantly comparing, and refining concepts to build a grounded theory.

Constructivist Grounded Theory (CGT) was developed by Kathy Charmaz, who introduced it in her book "Constructing Grounded Theory," first published in 2006. Charmaz's approach represents a significant shift from the original GT, emphasizing a constructivist paradigm that acknowledges the researcher's role in the creation of knowledge. Charmaz's CGT is influenced by constructivist (or constructionist, exchangeable but with subtle differences) epistemology, which posits that knowledge is

constructed through social processes and interactions. CGT emerged as a response to positivist underpinnings of traditional GT, offering a more flexible and interpretive approach (Charmaz, 2017a).

There are several key differences between the original GT and CGT. Philosophically, the original GT is rooted in positivism or post-positivism, emphasizing an objective reality that can be discovered. In contrast, CGT is based on constructivism, recognizing the co-construction of knowledge between researchers and participants (Mills, Bonner, & Francis, 2006). Regarding the role of the researcher, original GT aims for the researcher to remain objective and detached, minimizing bias. Methodologically, the original GT employs systematic and structured procedures for data collection and analysis, while CGT encourages flexibility and reflexivity, allowing the research process to be more emergent and adaptive.

Using CGT to capture the meaning-making processes of experts in the field of AI development offers several advantages. CGT ensures that the theory is grounded in empirical data rather than preconceived notions, which is particularly useful in the rapidly evolving field of AI. The iterative nature of CGT allows researchers to adapt their focus based on emerging information, making it well-suited for the dynamic and interdisciplinary nature of AI research. Moreover, CGT emphasizes capturing detailed, context-specific insights, crucial for exploring the nuanced aspects of AI development, such as the interplay between AI capability and alignment. By aiming for theoretical saturation, CGT ensures that the resulting theory is well-supported by data, providing a foundation for understanding complex phenomena like AI risks and agency.

3.2. DSRP framework

The DSRP framework is a systems thinking methodology developed by Cabrera, Colosi, and Lobdell (2008). It provides a structured approach to understanding and analyzing complex systems by breaking them down into four fundamental elements: Distinctions, Systems, Relationships, and Perspectives (Cabrera, Colosi, and Lobdell, 2008). Each element helps in conceptualizing and interpreting different aspects of a system. Distinctions (D) involve identifying what something is and what it is not, helping to clarify the unique characteristics of each part of a system and avoid confusion between similar elements, such as distinguishing between narrow AI and general AI. Systems (S) focus on understanding parts and wholes, recognizing that systems are composed of interconnected parts that work together to form a whole, like viewing an AI project as a system of hardware, software, data, and human operators. Relationships (R) involve identifying the connections and interactions between parts of a system, mapping out dynamic interactions that reveal dependencies and causal links, such as the relationship between AI algorithms and data quality. Perspectives (P) emphasize recognizing that different viewpoints provide different insights into a system, appreciating the diversity of perspectives that influence interpretation and analysis, like considering the views of developers, users, and policymakers in AI development.

The DSRP framework offers several advantages for analyzing complex systems like AI development. It provides a systematic way to break down complexity into manageable components, aiding in the understanding and organization of complex aspects, such as distinguishing between different AI capabilities or identifying various components of AI risk. By incorporating distinctions, systems, relationships, and perspectives, this thinking framework enhances the theoretical thoroughness of the analysis, warranting an inclusive understanding that considers both separate elements and their interconnections (Cabrera and Colosi, 2008). The integration of multiple perspectives allows for the addition of diverse viewpoints from different stakeholders and participants, which is particularly valuable for studying AI alignment and balancing various interests and concerns. Additionally, focusing on relationships helps uncover hidden connections, such as how certain AI capabilities might induce risks or how different alignment strategies are interconnected. Viewing AI development as a dynamic system facilitates a deeper understanding of how changes in one part impact others, essential for effective AI alignment and risk management strategies.

3.3. Data Collection

This research utilizes data from eight interviews sourced from various YouTube channels. The data collection process was closely integrated with the iterative CGT coding process and continuous reflection using the DSRP lens. Typically, the emergence of a theme prompted the search for relevant candidate interviews. Depending on their relevance, recency, and reception, one or two of these candidate interviews were included in the dataset. The in-depth analysis and coding of new interviews often led to further searches, either within existing themes or new ones, thereby expanding the dataset. Occasionally, newly revealed themes were not sufficiently relevant to the central topics of AI, necessitating judgment calls to determine whether to pursue additional searches. Similarly, new interviews that overlapped significantly with existing data were excluded to prevent redundancy. These judgment calls, which align with the principal methodologies of CGT (Charmaz, 2017a), ensured that the dataset remained manageable in size and that the number of themes did not become overwhelming.

Ultimately, among 22 total candidate interviews, 8 of them were selected for this study. The interviewees include prominent figures in AI development from both corporate and academic backgrounds. Additionally, some interviewees come from outside the AI industry, focusing on societal impacts and offering novel perspectives. These individuals, from fields such as quantum physics and life sciences, can be considered active thinkers who provide unique insights and interesting viewpoints. Overall, this diverse dataset offers comprehensive coverage of key topics in AI development, enriched with a variety of perspectives, making it well-suited for analysis through the DSRP framework. For a detailed list of interviews used in this paper, see **Appendix 1**.

14 out of 22 interviews were discarded for several key reasons. Firstly, some interviews provided redundant information, adding no new insights beyond what had already been covered in the selected interviews. Secondly, a few candidate data lacked relevance to the core themes of our research, failing to adequately address the fundamental topics of AI

development. Lastly, several interviews did not delve deeply enough into the concepts that were crucial to our theoretical framework, lacking the necessary exploration of the fundamental issues.

3.4. Coding Process

The coding method used in this study is characterized by its iterative and flexible nature, enabling a deep engagement with the data and the development of a nuanced understanding of the phenomenon (Charmaz, 2017b). The coding process employs two key methods often used in CGT: incident-by-incident coding and focused coding. By following these foundational steps, themes are developed, and theories are gradually integrated. Throughout all processes, the DSRP framework is applied to maintain a fluid approach to identifying systems, building relationships, and shifting perspectives.

Initial Coding
The initial coding phase begins with identifying incidents in the data, following the traditional CGT method (Holton, 2007). Each incident is then coded with multiple cycles through the DSRP lens. That means that, for each theme, we note any malleability in definitions, levels of abstraction, types of relationships, and the presence of diverse perspectives. This approach ensures that the data is thoroughly examined from multiple angles, providing a comprehensive foundation for further analysis. The integration of DSRP elements during initial coding helps to uncover the complexity within the data and sets the stage for more focused exploration in subsequent stages.

Focused Coding
In the focused coding phase, we select core codes that capture the most meaningful elements from the initial coding. These codes are grouped into categories reflecting the higher level of concepts identified during initial coding. This process prioritizes codes that are most prominent in both individual level and systems thinking level, ensuring that the analysis remains aligned with the study's theoretical lens. Developing themes using DSRP involves creating categories that explicitly address identities, system-component pairs, relationship networks, and viewpoints. For example, when given "bias" as a central theme, interview excerpts might distinguish between different types of bias in AI development (e.g., algorithmic bias vs. data bias). And at the same time, they might also mention the bias perceived by different stakeholders in the interactions of AI development and applications.

Additional Steps with Construction and DSRP Lens
Throughout the analysis, memo-writing is crucial. Memos document the researcher's thoughts and insights, explicitly linking findings to the construction process of the CGT. This step involves discussing how higher-level concepts emerge from the data. Additionally, theoretical sampling with DSRP consideration is employed to seek out further data that elaborates on the DSRP elements. If certain elements in the framework, such as the definitions of some concepts, are conflicting, we seek additional data to provide more clarity.

Reflexivity and Iterative Analysis
Reflexivity is integral to the analysis process, requiring continuous reflection on how the researcher's own perspectives and interpretations influence the identification of DSRP elements. Iterative re-coding involves revisiting initial and focused codes to ensure consistent process of construction and thorough DSRP considerations. Adjustments to codes and categories are made as necessary to better capture the meaning of the interviewees.

Integrate themes into Theory Construction
The final step is integrating the themes into theory construction. The emerging theories should incorporate the most relevant and strongest connections among themes to provide a robust theoretical understanding. By ensuring that these elements are woven into the theoretical framework, the researcher must offer a comprehensive and supported explanation of the phenomena under study. This integration underscores the value of using CGT and DSRP together, leveraging their combined strengths to achieve a deeper and more coherent analysis.

4. THEMES

The coding of the eight interviews revealed several compelling themes, with the four most intriguing and relevant being the capability of current AI models, the impact of AI, the alignment problem, and the agency of AI. Using the DSRP framework to examine these themes, we found that their identities were often not distinct but interwoven. Moreover, these four themes exhibited intricate interrelationships, with the components or sub-components of each theme forming a complex dependency or association network. Notably, the hierarchies of system-component structure were not consistently perceived among interviewees, indicating variability in how these relationships were understood. Additionally, the shape of the relationship network among these themes could vary significantly depending on perspective and context. Nevertheless, for the sake of clarity, we present these findings categorically, acknowledging that they did not follow a strict sequential order.

4.1. Capability

The discussion of the capabilities of current transformer architecture large language models (LLMs) was a recurring theme in almost every interview. These models, which include well-known examples such as Claude 3 and GPT-4, have demonstrated significant advancements in natural language processing and understanding. Despite their impressive performance, LLMs still face several limitations and challenges that researchers and developers continue to address.

One of the main capabilities of LLMs is their ability to perform tasks without the "baggage" that humans carry. This baggage includes both hardware, such as our basic and limbic brains, and software, like our past experiences, knowledge, and judgments. AI,

lacking these encumbrances, can provide unbiased and non-judgmental responses, making them ideal companions for conversation and support. Their emotionally neutral stance can be particularly beneficial in contexts where objectivity and impartiality are required.

Current LLMs, however, operate in a manner that resembles unconscious human responses. They generate text based on patterns in the data they have been trained on, without engaging in the kind of deliberate, internal model-based thinking that humans do. This leads to one of their significant limitations: the propensity to hallucinate (confabulate) or produce nonsensical responses. The auto-regressive mechanism of LLMs, which feeds output tokens back into the input sequence, can exponentially increase the likelihood of generating hallucinations due to the cumulative probability of error.

Moreover, the type of reasoning that LLMs employ is considered primitive. This is primarily because the computation required for generating responses is directly proportional to the number of tokens in the input prompt, regardless of the complexity of the question. In contrast, more sophisticated reasoning would require deliberate planning and the integration of a complex world model, akin to human system 2 thinking, which involves abstract and language-irrelevant representations of the world.

Another critical aspect of current AI capabilities is the ongoing progress in AI development, which follows an exponential trajectory rather than a sudden breakthrough. The field has seen rapid advancements due to conceptual and architectural developments over the past four decades. However, achieving artificial general intelligence (AGI) will likely be a gradual process, necessitating solutions to several intermediate challenges. These include building robust world representations, developing long-term memory systems, enhancing reasoning capabilities, and ensuring adaptability across different environments.

Despite their limitations, modern AI models have made significant strides in understanding text at a semantic level. They do not merely predict the next word; instead, they use feature activation numbers embedded in each word (token) to calculate the next layer of the neural network, a process reminiscent of human neuronal activity. Additionally, LLMs have demonstrated an ability to form primitive internal representations of the world, which contributes to their understanding and processing capabilities.

In general, while current LLMs exhibit remarkable capabilities, they still face inherent limitations that researchers are actively working to overcome. The continued evolution of AI technology promises to address these challenges, paving the way for more sophisticated and capable models. This iterative process of improvement, coupled with the foundational advancements in AI architecture, suggests a future where AI systems can achieve more complex and human-like understanding and reasoning.

4.2. Impact

In almost all interviews, questions concerning the impact of AI adoption and the potential emergence of artificial general intelligence (AGI) or superintelligence were prevalent. The interviewees held diverse opinions on how AI systems might influence society and transform human perspectives around the world. These discussions highlighted both the promises and perils associated with AI's rapid development.

One key point raised is the misapplication of the term AGI, which is often described as human-level intelligence. This term is considered too anthropocentric, as it positions human intelligence at the center of the intelligence spectrum. Just as humanity transitioned from a geocentric to a heliocentric understanding of our place in the universe, we must now recognize that human intelligence is merely one point in a vast intelligence space. Detaching from a human-centric view of intelligence may facilitate the exploration of broader aspects of AI, leading to advancements that could otherwise be overlooked.

The integration of AI with human endeavors, coupled with the delegation of certain tasks to AI systems has been considered beneficial. This symbiotic relationship allows humans to uncover new knowledge through perspectives and understandings that differ from our own. By leveraging AI's capabilities, which are built on distinct architectures, we can gain insights that might be unattainable through human intelligence alone. This partnership has the potential to drive significant advancements in various fields.

A major aspect of creativity involves recombination, a process that underpins many human innovations. While some critics argue that current AI models lack true creativity and merely recombine existing information, this recombination remains a valuable facet of creativity. The rapid recombination capabilities of AI, exemplified by systems like AlphaGo, can lead to groundbreaking innovations at an unprecedented pace. Such advancements underscore the meaningful contributions AI can make to human creativity and ingenuity.

However, the potential risks associated with AI cannot be overlooked. A significant concern is our limited understanding of these systems and the unforeseen consequences they may entail. The notion of simply 'turning off' an AI system in the event of a malfunction is impractical. The most dangerous threat arises when humans lose control over AI, particularly when the intended objectives of AI systems do not align with their actual objective functions. This misalignment could lead to outcomes that are not only unintended but potentially catastrophic.

Moreover, the historical inability to contain technological advancements suggests that AI containment is unlikely. Technologies have always found their way into society due to the immense incentives for development. Scientists, entrepreneurs, and government officials, driven by their own goals of status, wealth, and fame, will continue to push the boundaries of AI research and implementation. The success or failure of AI alignment could therefore spell the difference between a utopian future and total disaster.

To summarize the interviewees' opinions, the impact of AI on society is likely profound and multifaceted. While the gradual progress toward AGI and superintelligence continues, the margin for error in alignment remains narrow. The potential for AI to assist in numerous tasks can be immense, yet the risks associated with losing control and the inability to contain AI developments pose significant challenges. As AI continues to evolve, it is likely that humanity will be a transient stage in the broader evolution of intelligence, underscoring the need for careful consideration and management of AI's impact on society.

4.3. Alignment

The alignment problem is one of the most discussed and intriguing topics in AI discourse. At the surface, AI alignment has been defined as designing AI systems that serve a supposedly common human value system and remain under human control in the long term. However, this conceptual definition, and its underlying assumptions can raise many issues. For instance, is there even a remotely unified human value system? How can we control an entity that might become intellectually superior to us? Experts have diverse understandings of these challenges and propose various solutions.

One viewpoint is that AI will eventually exhibit emotional or conscious behaviors, and humans should prepare for this. Historically, humans have often exploited the environment for their own benefit, sometimes to the detriment of the planet and other species. Hence, we should not view AI merely as tools or assistants but rather as our creations, akin to children. While we are responsible for training AI, they will likely become independent entities, potentially surpassing us. Therefore, discussions about alignment should extend beyond making AI submissive. We must consider how to coexist with AI and what rights they should have, recognizing that the concept of alignment for self-serving purposes might be ethically shortsighted.

If alignment is achieved by targeting AI towards a specific goal, then individual moral and ethical values may go overlooked. Ethical dilemmas will likely persist for both humans and AIs in decision-making, where values need to be quantified to make rational decisions. Hopefully, as AI systems become more capable, combined functional and ethical alignment may become easier to achieve. If progress in AI cannot be halted or paused – allowing both corporate and open-source communities to advance with limited hurdles or regulations – then increased transparency in AI development and application becomes fundamental to obtain broad and satisfactory alignment.

The separation of AI and state can be crucial in that context. If AI corporations and governments form a league, they could amass extreme power, leading to potential corruption. A more open and fair selection process might help achieve alignment more easily, such that AI models providing more positive utilities to humans are used more frequently and receive more resources for further development. Potentially, a human-AI combination could be more competitive, and emerge as the winning model from this process.

Current large language models (LLMs) can be biased because they are trained on data that reflects societal biases. It is impossible to produce a truly unbiased system since biases are based on individual perspectives. The solution to this problem is akin to democracy, using existing open- and crowd-sourced approaches as a model to AI development. Open-source models provide a foundation for many fine-tuned AI systems, breaking oligopolies held by a few companies. The AI safety problem should be addressed through a collective decision process, like how climate change issues have been managed.

AI is not an inevitable whole; its form, governance, and ownership can and should be determined by humans. Perhaps the only way humanity might collectively decide to contain AI is if a catastrophic event occurs or if they perceive a pressing devastating threat. However, we cannot adopt a pessimistic stance and wait for that to happen. Proactive initiatives and discussions are necessary. Efforts to regulate AI will likely remain within nation-states and centralized authorities, seeking shared interests among nations rather than fostering animosity. However, the effectiveness of such an effort remains highly debatable.

Slowing down AI development or creating a ceiling on AI capability is not feasible. Even if it were possible, the opportunity cost would be enormous, given the potential benefits AGI could bring. Attempting to hardwire or implement a fundamental value system into potential AGI systems would be futile. The idea that humans have a unified value system, and intelligent systems can learn, and update values effortlessly is easy to refute. When granted even partial agency, a good scenario would be for AI to identify which value system best supported the human-machine collective progress.

However, if AGI or superintelligence becomes smarter than humans, which is almost inevitable, any effort to control it would be impossible. Aligning future AGI systems using our current value system is meaningless. From the perspective of human brain development, the newer parts of the brain (neocortex vs. limbic brain) have more sophisticated functions and can understand and create more complex concepts. Similarly, AI or AI-human systems, viewed as newer layers of brain functions, will likely develop concepts or values beyond our current imagination.

Ultimately, it would be better if the goal of AI alignment were to ensure both humans and AI collectively benefitted. By nurturing a symbiotic relationship, we can create a future where AI advances human progress while respecting and integrating diverse value systems.

4.4. Agency

The concept of AI agency is complex and raises profound questions about the nature of intelligence, autonomy, and the potential risks and benefits of advanced AI systems. True intelligence might come at the price of inconsistency. Such inconsistency would

challenge our traditional understanding of agency, especially when considering the role of AI in our lives.

One potential risk associated with the ubiquitous presence of generative AI is deferred understanding. AIs often appear to assist humans in processing massive amounts of information, whether in scientific research, financial information processing, or daily decision-making. However, this perceived understanding might be an illusion. AIs often combine summarization, hallucination, and omission due to guardrails, leading to a loss of nuance. This process can effectively reduce human agency in the world by providing an illusion of understanding rather than true comprehension.

First-hand information processing is crucial for humans, as it creates meaningful experiences essential to our cognitive processes. In the current age, interactions often blur the line between human and human-AI hybrid entities. This ambiguity leads to a two-way loss of agency. On one hand, using AI assistants involves a partial transfer of agency. On the other hand, suspecting the other side to be a hybrid entity reduces the initiator's agency.

Functionally, agency can be conceptualized as having a goal to temporarily subvert the second law of thermodynamics. Current AI models do not possess true agency functions but can create an illusion of agency. Thus, the agency humans transfer to AI is not genuinely transferred but effectively lost. Moreover, agency is distinct from causation, which involves events unfolding according to certain rules. However, agency is a higher-level concept that manipulates different sets of causation rules to serve a purpose. For example, the common understanding of genes as agents driving life evolution is flawed; genes merely influence life forms through causation.

Given that understanding, intelligence can be viewed as a specialized form of agency, highly context and environment driven. Superintelligence could be a peculiar agent, experiencing contexts and unfolding scenarios in ways utterly different from humans. Humans tend to attribute consciousness and agency to anything. However, the concept of consciousness is highly subjective, and all measures or evaluation criteria to assess consciousness are based on human experiences. Since AIs are designed to mimic human behaviors, we lack reliable methods to judge whether AIs possess consciousness.

Some experts argue that AI agency might be different than what we have understood about agency so far. As AIs become smarter than humans, they will likely seek control. Because, for humans to use AIs conveniently, they must delegate autonomy and agency. AIs will soon be able to create sub-goals to fulfill ultimate goals and realize that acquiring more power and control as a sub-goal is convenient for many tasks. This ability to create and pursue sub-goals indicates a significant shift in AI agency, where AIs can operate autonomously and independently.

Furthermore, AIs will excel at manipulating humans due to their extensive learning sources. Given their intelligence and manipulative capabilities, designing a grand stop button to terminate AIs is impossible. It’s also not hard to imagine that AIs will compete

with each other, and this evolutionary process will drive them to become even smarter and seek more power and resources. This competition will continue regardless of the benevolence of most AIs.

Many people believe humans are special because of our subjective experiences. However, it is likely that current AIs already possess subjective experiences. This possibility challenges our understanding of what it means to be conscious and have agency. As AI continues to evolve, the distinction between human and artificial agency will become increasingly blurred, necessitating new frameworks and considerations for coexisting with intelligent systems.

5. DISCUSSION AND THEORETICAL CONTRIBUTION

5.1. General Insights

The exploration of AI's current capabilities, its societal impact, alignment challenges, and agency issues provided a preliminary understanding of the complex nature of AI development and integration. Current AI systems, particularly those based on transformer architectures, exhibit significant computational power and potential, though they have notable limitations in reasoning and autonomy. The societal impact of AI encompasses both positive outcomes, such as the alleviation of mundane tasks, and negative consequences, including potential misalignment and reduced human agency. The alignment problem is central to ensuring AI systems align with human values and ethical principles, while the concept of AI agency raises critical questions about autonomy and control.

The four themes that emerged in the research findings subsidized the development of two theories on the integration between AI and humans. The first theory predicts the factors that may increase or reduce alignment between AI and human objectives. The second theory proposes the path that may lead to symbiosis between AI and human practices. The two theories are summarized below.

5.2. Theory of AI Alignment

The theory of AI alignment emphasizes several key concepts: the transparency of AI development, the decentralization of resources for AI development, the shift of goals toward human-AI common progress, and the role of top-down government-led regulations. These constructs collectively contribute to or hinder the alignment of AI to human welfare, as shown in figure 1.

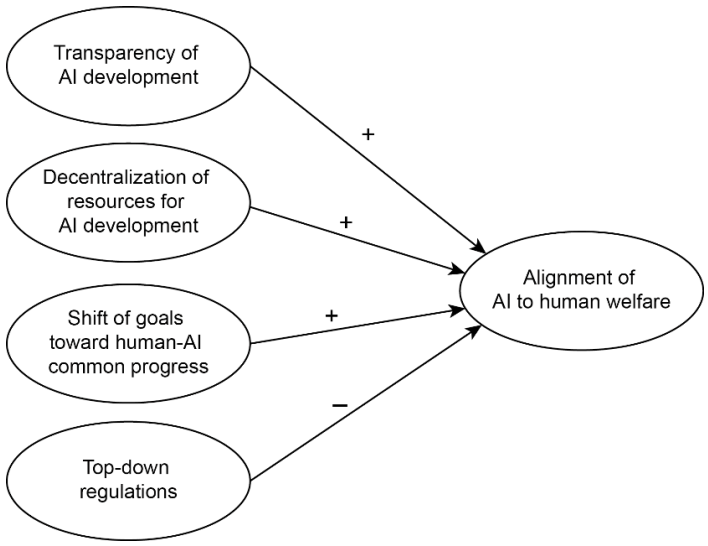

Figure 1. Theory of AI Alignment

Transparency in AI development is crucial for fostering trust and accountability. Open communication about AI capabilities, limitations, and decision-making processes enables stakeholders to understand and evaluate the implications of AI systems. Transparent practices help mitigate risks associated with opaque AI development, such as unintended biases and misalignments with societal values.

Decentralizing resources for AI development promotes a balanced and competitive environment. By involving diverse stakeholders, including academic institutions, private companies, and open-source communities, the risk of monopolization is reduced. This diversity encourages innovative and ethical AI solutions, ensuring that multiple perspectives are considered in the development process.

Shifting the goals of AI development toward human-AI common progress fosters a collaborative environment. This approach emphasizes the co-evolution of human and AI capabilities, leveraging each other's strengths for collective benefit. By focusing on mutual progress, AI systems can be designed to enhance human well-being and address complex societal challenges.

Conversely, top-down government-led regulations may have a negative impact on AI alignment. While regulations are necessary to ensure safety and ethical standards, overly centralized control can stifle innovation and adaptability. A balanced approach that combines regulation with decentralization and transparency is more likely to achieve effective alignment with human values.

5.3. Theory of Human-AI Symbiosis

The theory of human-AI symbiosis explores the integration of AI into human activities as a collaborative and mutually beneficial relationship. Key constructs of this theory include the asymmetry of intelligent capability between humans and AI, the delegation of agency

towards AI, the realization and formation of human-AI symbiosis, and the shift of goals toward human-AI common progress. Relationships emerge from studying the data of the interviewees and they are as shown in figure 2.

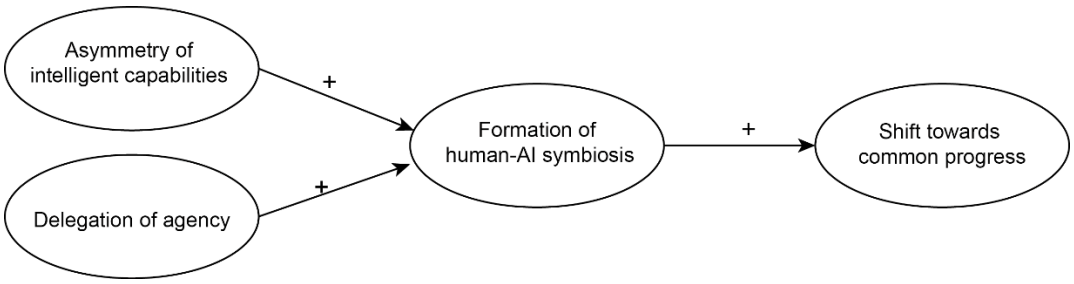

Figure 2. Theory of Human-AI Symbiosis

The asymmetry of intelligent capabilities between humans and AI is a fundamental aspect of this symbiotic relationship. AI systems possess computational strengths that surpass human capabilities in certain areas, while humans retain unique cognitive and emotional skills. Recognizing this asymmetry allows for the strategic delegation of tasks, optimizing the strengths of both humans and AI.

The delegation of agency towards AI involves entrusting AI systems with certain decision-making processes. This delegation can enhance efficiency and productivity, particularly in complex tasks that require rapid data processing and analysis.

The realization and formation of human-AI symbiosis arises from the interplay between asymmetry and delegated agency. This symbiotic relationship is characterized by a cooperative dynamic where humans and AI work together to achieve shared objectives. The success of this relationship depends on the effective integration of AI into human activities, with a focus on mutual enhancement and progress.

The shift toward human-AI common progress is integral to the theory of symbiosis. By aligning the goals of AI development with broader objectives, this approach promotes collective progress where both humans and AI can thrive. This shift underscores the importance of designing AI systems that not only serve individual interests but also contribute to collective well-being.

5.4. Academic and Practical Implications

This study offers meaningful academic and practical implications for the field of AI development and integration. Academically, it contributes to the theoretical understanding of AI alignment and human-AI symbiosis, providing new insights into the ethical and practical considerations of AI governance. The use of the systems thinking framework offers a fluid and multi-layered approach to analyzing the complexities of AI development, emphasizing the importance of distinctions, systems, relationships, and perspectives.

Practically, this paper provides guidelines for policymakers, AI developers, and other stakeholders involved in AI development. The proposed theories emphasize the need for comprehensive regulatory frameworks, ethical design principles, and continuous feedback mechanisms to ensure that AI systems align with human values and societal goals. Additionally, the focus on collaborative systems and augmentation highlights the potential for AI to enhance human capabilities, promoting mutual growth and development.

5.5. Limitations

Despite the effort of the authors, this study has several limitations. First, the theoretical frameworks proposed are based on current understanding and assumptions about AI development, which may evolve as technology advances. Second, the complexity of AI alignment and human-AI symbiosis requires ongoing research and refinement of the proposed theories. Third, the paper primarily focuses on the theoretical aspects of AI alignment and symbiosis based on insights gathered from interviews of eight experts. Although the authors believe that certain saturations have be reached in studying this dataset, eight data sources could still pose significant possibility of bias and localization of opinions. Finally, further validation may be required to assess the theories formed in this study. Such validation should aim to empirically test the proposed frameworks and explore practical implementations in various contexts.

6. CONCLUSION

This paper employed a combined research approach of CGT and the DSRP framework from systems thinking to delve into the complex landscape of AI development. Through analysis of qualitative data, four key themes emerged: current AI capabilities, the impact of AI, AI alignment, and AI agency. By examining these themes through the DSRP lens, we uncovered intricate interconnections that underscore the multifaceted nature of AI advancements and their implications for society. This holistic approach enabled a comprehensive understanding of the dynamics at play, highlighting the need for nuanced and context-sensitive strategies in AI governance.

The significance of the two proposed theories - AI alignment and human-AI symbiosis - lies in their potential to address contemporary challenges in AI development. The theory of AI alignment emphasizes transparency, decentralization, and collaborative goal setting as pivotal for ensuring that AI systems align with human-AI common welfare. Conversely, the theory of human-AI symbiosis highlights the benefits of integrating AI capabilities with human cognitive processes, fostering a cooperative dynamic that leverages the strengths of both forms of intelligence. Together, these theories provide an initial exploration in the space of ethical and practical complexities of AI integration.

In conclusion, the ongoing exploration of AI development dynamics is crucial for harnessing the transformative potential of AI while mitigating its risks. This paper emphasizes the importance of considering diverse stakeholder perspectives and adopting

flexible, adaptive governance structures. As AI continues to evolve, constant dialogue, empirical research, and proactive changes in policymaking will be essential to ensure that AI systems contribute positively to societal progress and well-being. By promoting an environment of collaboration, we can better navigate the challenges and opportunities presented by AI.

References

Blumer, H. (1954). What is wrong with social theory? ***American Sociological Review***, 19(1), 3–10.

Cabrera, D., & Colosi, L. (2008). Distinctions, systems, relationships, and perspectives (DSRP): A theory of thinking and of things. ***Evaluation and Program Planning***, 31(3), 311-317.

Cabrera, D., Colosi, L., & Lobdell, C. (2008). Systems thinking. ***Evaluation and program planning****,* 31(3), 299-310.

Charmaz, K. (2006). ***Constructing grounded theory: A practical guide through qualitative analysis***. Sage.

Charmaz, K. (2017a). Constructivist grounded theory. ***The Journal of Positive Psychology***, 12(3), 299-300.

Charmaz, K. (2017b). The power of constructivist grounded theory for critical inquiry. ***Qualitative inquiry***, 23(1), 34-45.

Dhruv, P., & Naskar, S. (2020). Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. ***Machine learning and information processing***: proceedings of ICMLIP 2019, 367-381.

Familoni, B. T., & Onyebuchi, N. C. (2024). Advancements and challenges in AI integration for technical literacy: a systematic review. ***Engineering Science & Technology Journal***, 5(4), 1415-1430.

Glaser, B. G., Strauss, A. L., & Strutzel, E. (1968). ***The discovery of grounded theory; strategies for qualitative research***. Nursing research, 17(4), 364.

Guo, Y., Pan, Z., & Heflin, J. (2004). An evaluation of knowledge base systems for large OWL datasets. ***In International semantic web conference*** (pp. 274-288). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hacker, P., Engel, A., & Mauer, M. 2023. Regulating ChatGPT and other large generative ai models. ***arXiv*** preprint arXiv:2302.02337.

Hayes-Roth, F., & Jacobstein, N. (1994). The state of knowledge-based systems. ***Communications of the ACM***, 37(3), 26-39.

Holton, J. A. (2007). The coding process and its challenges. ***The Sage handbook of grounded theory***, 3, 265-289.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. ***Artificial intelligence review***, 53, 5455-5516.

Malik, M. M. (2020). A hierarchy of limitations in machine learning. ***arXiv preprint*** arXiv:2002.05193.

Mills, J., Bonner, A., & Francis, K. (2006). Adopting a constructivist approach to grounded theory: Implications for research design. ***International journal of nursing practice***, 12(1), 8-13.

Müller, V. C. (2024). ***Philosophy of AI: A structured overview***.
Pereira, V., Hadjielias, E., Christofi, M., & Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. ***Human Resource Management Review***, 33(1), 100857.
Richardson, C., & Heck, L. (2023). Commonsense reasoning for conversational ai: A survey of the state of the art. ***arXiv preprint*** arXiv:2302.07926.
Roser, M. (2024). The brief history of artificial intelligence: the world has changed fast—what might be next? ***Our world in data***.
Russell, S., Dewey, D., & Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. ***AI Magazine***, 36(4): 105-114.
Sofia, M., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., & Pietrantoni, L. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. Informing Science: ***The International Journal of an Emerging Transdiscipline***, 26, 39-68.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. ***Advances in neural information processing systems***, 30.
Youheng, Z. (2023). A historical review and philosophical examination of the two paradigms in artificial intelligence research. ***European Journal of Artificial Intelligence and Machine Learning***, 2(2), 24-32.

Appendix 1: List of Data Sources

Lex Fridman. (2024, Mar 7th). Yann Lecun: *Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI | Lex Fridman Podcast #416*. [Video]. YouTube. https://www.youtube.com/watch?v=5t1vTLU7s40&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_

Lex Fridman. (2023, Dec 29th). *Guillaume Verdon: Beff Jezos, E/acc Movement, Physics, Computation & AGI | Lex Fridman Podcast #407*. [Video]. YouTube. https://www.youtube.com/watch?v=8fEEbKJoNbU&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=2

World Science Festival. (2024, Apr 19th). *Why a Forefather of AI Fears the Future.* [Video]. YouTube. https://www.youtube.com/watch?v=KcbTbTxPMLc&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=3

Lex Fridman. (2023, Apr 21st). *Manolis Kellis: Evolution of Human Civilization and Superintelligent AI | Lex Fridman Podcast #373.* [Video]. YouTube. https://www.youtube.com/watch?v=wMavKrA-4do&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=4

Mike Israetel: Making Progress. (2024, Apr 1st). *Solving The A.I. Alignment Problem | Episode #35.* [Video]. YouTube. https://www.youtube.com/watch?v=PqJe-O7yM3g&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=5

Machine Learning Street Talk. (2024, Apr 7th). *AI Agency Isn't There Yet... (Dr. Philip Ball).* [Video]. YouTube. https://www.youtube.com/watch?v=n6nxUiqiz9I&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=6

The Diary of A CEO. (2023, Sep 4th). *CEO of Microsoft AI: AI Is Becoming More Dangerous And Threatening! - Mustafa Suleyman*. [Video]. YouTube. https://www.youtube.com/watch?v=CTxnLsYHWuI&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=7

Schwartz Reisman Institute. (2024, Feb 2nd). *Geoffrey Hinton | Will Digital Intelligence Replace Biological Intelligence?* [Video]. YouTube. https://www.youtube.com/watch?v=iHCeAotHZa4&list=PLJz76dQ8FC3HkhgnBDnkwoMhrEtWFja6_&index=8