Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Teologiczny

Maciej Mróz

Morality of Artificial Agents

Moralność sztucznych agentów

ROZPRAWA DOKTORSKA

DZIEDZINA NAUK TEOLOGICZNYCH
DYSCYPLINA NAUKI TEOLOGICZNE

Praca doktorska napisana pod kierunkiem:
prof. dr. hab. Krzysztofa Stachewicza

Poznań, 2025r.

Table of Contents

Introduction

The rapid development and proliferation of artificial intelligence (AI) demand a diverse, multi-voiced debate about how to shape technological progress. AI's potential to reshape social and economic reality is arguably unprecedented for any technology in human history. Therefore, this debate must be not only pluralistic and inclusive in nature but also well informed; it should be conceptually rigorous, technically literate, and attentive to social contexts. This dissertation aims to contribute to the scholarly discourse in AI ethics while remaining grounded in a theological perspective.

The context of artificial intelligence calls for multidisciplinary approaches. As will be shown throughout this thesis, discourse on AI is often clouded by terminological instability and divergent uses of key terms across disciplines, including computer science and philosophy. A clear example is the term "value" itself. Central to this thesis, it denotes two distinct notions: within moral theory, a claim about the good; within computing, a parameter to be set. Building on this framing, the question may be asked: what does it mean for an algorithm to be good? To address this issue, this study offers two contributions: a critical review of current debates on the morality of artificial agents and an exploration of further dimensions of this already complex issue.

The first chapter serves as an orientation within the complex mosaic of the current AI landscape. It sets the stage for further discussion by providing an overview of the evolution of AI definitions and summarizing key concepts in the field of AI ethics. The second chapter reviews the notion of artificial agency in its many forms. This includes a presentation of diverse perspectives throughout the development of philosophical concepts on agency in general and artificial agency in particular. It also explains why in the context of this dissertation speaking of *artificial agency* is preferred over *artificial intelligence*. The third chapter focuses on the notion of equipping artificial agents with capacity for moral deliberation. This includes concepts of machine ethics and Artificial Moral Agents (AMAs) and reviews the academic discourse on those subjects to date. Chapter four seeks to deliberate whether morality indeed can be computed. This presents philosophical perspectives from both secular and Christian traditions. On the top of challenges with ascribing artificial agents with responsibility, it introduces to the debate also the new dimension: the notion of moral act as a creative act and its implications for the concept of machine ethics. The analysis also critically reviews some attempts at building computational

morality exemplifying the problem of too narrow and reductionistic understanding of philosophical concepts by some researchers in the field of robotics and artificial intelligence. Finally, the study distills conclusions from the contemporary debate on artificial moral agency and maps their implications for AI ethics and the wider social context.

All the parts of the thesis operate in the field of philosophy; however, they are being supplemented by the relevant discourse from the point of view of AI theorists and researchers. This thesis strongly claims that the disconnect between these two disciplines is the root cause of many challenges in the realm of AI development. Bridging these two worlds can be the key task for future research while these systems are growing more in complexity. Especially, anthropomorphizing machines leads often to conflated understandings impacting the very foundations of AI research directions. Imposing on machines the terms used for millennia to describe human properties has profound implications especially in the ethical and social context. On the other hand, some concepts drawn from computer science can contribute to the development of philosophy as it has been already demonstrated by some philosophers[1]. Moreover, technically literate deliberation on AI related phenomena posits a great opportunity for anthropology, because their analysis leads consequentially to profound questions about the nature and destiny of human.

An important feature of this dissertation is its grounding in the discipline of theology. From a theological perspective, technologies are human artifacts that are part of culture. This way they are becoming new "places" of theology. This applies to many issues of technology in general and its various applications. However, in this landscape of technical artifacts, artificial intelligence is a special case, due to its nature and transformative potential. Unlike any other artifacts, AI is built with the intention of imitating and even replacing human actions as much as possible. Technological artifacts that are part of culture can be treated as one of the new "sites" of man, and thus also as theological places. Reflection on algorithms created to imitate and replace intelligent human actions leads to questions about the nature of man, his functioning, and his destiny being asked in a new context (Mróz, 2024). Keeping the proper proportions in mind, and remembering that for theology, its main place and source is the Revelation, it should be noted that other phenomena, including various manifestations of human activity and reflection, also belong to theological places (loci theologici). Although according to various classifications, including the commonly cited one by Melchior Cano, the *locus theologicus* related to culture should be placed at the bottom of the hierarchy of

[1] The great example is the work of professor Luciano Floridi, who introduced number of new philosophical methods inspired by computer science concepts, such as Floridi's "Levels of Abstraction" method.

importance, it should be noted that it is becoming the subject of growing interest in the discipline of theological sciences (Mróz, 2024). To fully understand the message of salvation resulting from Christian Revelation, one must also fully understand who man is. This is one of the core claims of Karl Rahner's anthropological method (Dzidek and Sikora, 2018, p. 155). In this context, viewing culture as a manifestation of human activity is, as Kulisz argues, integral to theology's identity. This identity requires interpreting and communicating the content of Revelation using the context of contemporary human life. It leads to new tasks for theology, which assume serving faith and culture, because in culture and through culture human beings gain true and full humanity. It is also the place where they question their ultimate destiny. Kulisz refers in this context to the teaching of the Church contained in the "Pastoral Constitution on the Church in the Modern World *Gaudium et spes*" (Kulisz 2012, p. 254). Benanti speaks about a phenomenon that he calls the *techno-human condition* of human nature (Benanti, 2016). He refers to the term techno-human condition as the way in which human beings have always experienced and understood their existence: through engagement with the world mediated by tools and technological artifacts (Benanti, 2023). The significance of this fact for theology is highlighted also by one more thing: not only do humans create culture, including technological artifacts, but these artifacts also shape their existence in the world. Sir Winston Churchill aptly expressed this idea in his speech to the House of Lords in 1943, saying: "We shape our buildings; thereafter they shape us" (Churchill, 1943). For these reasons, AI is becoming an object of increasing interest in theology. Just to name the key efforts in this context, is worth calling out numerous remarks by Pope Francis on AI, the Pontifical Academy for Life's "Rome Call for AI Ethics"(2020), and "Antiqua et nova" (2025), a doctrinal note of the Catholic Church jointly issued by the Dicastery for the Doctrine of the Faith and the Dicastery for Culture and Education. "Antiqua et nova" addresses the relationship between artificial intelligence and human intelligence and offers reflections on the anthropological and ethical challenges raised by AI. Also, in explaining his choice of pontifical name, Pope Leo XIV referenced AI, highlighting the new challenges it poses and its transformative power, potentially comparable in impact to the Industrial Revolution (Davis, 2025).

As far as research methodology is concerned this investigation is exegetical and critical: it first reconstructs the arguments and then evaluate them. The analysis proceeds from interpretation to evaluation: historical context informs the readings, while contemporary analytic tools – logical reconstruction, counterexamples, and reflective equilibrium – guide the assessment. The project is comparative and problem-driven rather

than author-centered. Wherever justified it offers historically sensitive readings, drawing from philosophical tradition, before turning to systematic assessment in contemporary terms. The aim is not exhaustive exegesis but targeted analysis of arguments bearing on morality of artificial agents. Texts outside this scope are noted only where necessary. With regard to the above, this research aims to: (1) review the debate about the possibility, rationale, and conditions for creating artificial moral agents; (2) analyze the context in which this debate arises; and (3) identify the implications of the main approaches. This work uses elements of phenomenological method to more fully grasp both the nature of artificial agents and the essence of morality and its associated values (e.g., intentionality, agency, responsibility, normativity). It also leverages elements of source analysis and hermeneutics, by a systematic review of literature on machine ethics and Artificial Moral Agents (AMAs), critical interpretation of key texts, and application of the resulting conclusions to the project's focus and contemporary implementations. Finally, it conducts analysis and synthesis—decomposing concepts, comparing models and arguments, and integrating the findings into a coherent conceptual framework and a set of criteria for assessing artificial agents' capacity for moral deliberation.

The main challenge in writing this thesis was a rapid development in the field of AI in the recent time and associated with it the development of AI ethics. Literally thousands of academic papers and books have been published concerning ethical dimensions of artificial intelligence in past six years. Keeping up with the debate poses significant challenge itself. That said, the critical literature review demonstrated that the vast amount of these publications presents limited novelty, often being new reframing of already known and widely discussed concepts, rarely offering true breakthroughs. Another challenge arises from the inherently multidimensional nature of AI phenomena. This dissertation deliberately focuses on the question of whether AI-like artificial agents can be regarded as moral agents and, if so, in what ways. Naturally, this line of inquiry requires engagement with a range of related topics in order to establish the requisite conceptual groundwork, although such discussions are treated only in a concise manner, for providing necessary context. Accordingly, questions about the moral status of artificial entities, although interesting and closely related, fall outside the scope of this work and are discussed only briefly. Similarly, questions about the moral dimensions of shared or collective forms of artificial agency, such as states, institutions, and corporations, are intentionally set aside.

In the process of writing this dissertation Generative AI tools have been used for proofreading, grammar correction and literature gap analysis. For this purposes, the

following AI models have been used: GPT-4o, Gemini 2.5 Pro, Claude Opus 4.1. That said, this dissertation presents original work by its author. Generative AI has only been used in accordance with the UAM guidelines. For more information on the extent and nature of AI usage, please contact the author.

Chapter 1.
Navigating the Complex Mosaic of Artificial Intelligence

Artificial intelligence is a multidimensional and multidisciplinary phenomenon. There is no single, widely accepted definition of AI. Yet the definition one adopts, explicitly or implicitly, has far-reaching consequences: it shapes how AI systems are perceived and developed. This chapter therefore surveys various approaches to defining AI and traces how those definitions have evolved across contexts. A crucial fact is that in principle these systems are designed and developed to mimic and replace human cognitive abilities. This invites the anthropomorphizing of machines and leads to significant consequences that extend beyond technology itself. Systems with such autonomy and impact have enormous potential to shape the human world and the natural environment. The scale of AI's transformative power is arguably unprecedented in the history of technology. This calls for viewing AI as more than technology and for treating it as a *sociotechnical* phenomenon. Therefore, the challenges posed by the functioning of AI systems must be addressed through regulation and ethics. The discipline of AI ethics has emerged as a response to these needs. In the final part of this chapter, AI ethics, in its diverse manifestations, will be reviewed to situate the analysis that follows. Some of those issues may at first seem less relevant and less aligned with the research goal of this dissertation, but in the final chapters they will be leveraged and linked to the core reflection. Therefore, all the elements discussed in this chapter are foundational for understanding the context in which the questions about morality of artificial agents are raised.

1.1 Definitions of Artificial Intelligence

1.1.1. Dictionary definition of AI

The Oxford English Dictionary characterizes artificial intelligence as "*The capacity of computers or other machines to exhibit or simulate intelligent behavior; the field of study concerned with this. In later use also: software used to perform tasks or produce output previously thought to require human intelligence, esp. by using machine learning to extrapolate from large collections of data*" (Oxford English Dictionary, 2023). This framework emphasizes the fundamentally technological nature of artificial intelligence as systems explicitly designed and constructed to replicate, simulate, or potentially replace

human cognitive functions. The focus on imitating and replacing intelligent human actions reveals the anthropocentric orientation that has characterized the discourse on artificial intelligence since its inception. However, this technological approach immediately raises philosophical questions about the nature of intelligence itself. Speaking about machines "imitating" human cognitive functions, implicitly assumes that human cognitive functions can be adequately characterized in functional terms and that these functions *can* be implemented in non-biological substrates. The dictionary definition thus contains important metaphysical assumptions about the mind, intelligence, and the relationship between biological and artificial systems: assumptions that deserve careful philosophical analysis.

1.1.2. The origin and marketing dimension of the term

John McCarthy coined the phrase *Artificial Intelligence* in 1955 while preparing a proposal for the 1956 Dartmouth Summer Research Project, a conference that formally inaugurated AI as a distinct field of research. McCarthy defined Artificial Intelligence as "*the science and engineering of making intelligent machines*" (McCarthy et al., 2006). McCarthy's choice was neither neutral nor purely descriptive; he later admitted that he chose the term "artificial intelligence" in part to distinguish his proposed research program from the already established fields of cybernetics and automata theory, and in this way to attract both funding and talent to this initiative (Kline, 2011). The genealogy of the term *artificial intelligence* reveals its partly strategic and commercial origins, offering key insights into how marketing considerations have shaped scientific discourse. The marketing dimension of this fundamental moment cannot be overstated: by choosing language that suggested the possibility of creating true intelligence, rather than just computational tools, McCarthy established a narrative framework that has profoundly influenced both public perception and research directions for decades to come. The success of this term in attracting attention and resources shows how scientific fields can be shaped by rhetorical and strategic considerations, not just purely epistemic ones.

1.1.3. Russell and Norvig's Definitional Framework

Stuart Russell and Peter Norvig, in their textbook "Artificial Intelligence: A Modern Approach" present a systematic taxonomy of definitions of artificial intelligence that has become canonical in the field. Their framework organizes definitions along two orthogonal dimensions: whether the system aims to match human performance or achieve ideal

rationality, and whether success is measured by internal processes (thinking) or external behavior (action). This gives four different approaches to defining artificial intelligence:

- systems that think like humans (cognitive modeling approach)
- think rationally (thinking laws approach)
- act like humans (Turing test approach)
- and act rationally (rational agent approach)

Russell and Norvig advocate the rational agent approach, defining artificial intelligence as "the study of agents that receive stimuli from their environment and perform actions" in order to achieve the best outcome or, under conditions of uncertainty, the best expected outcome (Russell and Norvig, 2021, p. 26). This definition shifts the focus from anthropomorphic comparisons to criteria of optimality, suggesting that artificial intelligence does not need to mimic human intelligence, but rather achieve effective performance in complex environments. The rational agent framework has proven particularly influential in contemporary AI research, providing a mathematical basis for analyzing intelligent behavior without the need to directly compare it to human cognition.

The philosophical implications of this choice of definition are significant. By prioritizing rationality over human-like qualities, Russell and Norvig indirectly endorse a functionalist view of intelligence; one that defines mental states by their causal role rather than by physical instance or phenomenological characteristics. This approach avoids difficult questions about consciousness and subjective experience, while maintaining that true intelligence can be realized in artificial systems. However, this raises new questions: whose conception of rationality serves as the standard? How should we decide between different models of rational behavior?

1.1.4. Narrow Vs. General Artificial Intelligence

The distinction between narrow and artificial general intelligence is useful one when it comes to separating current technological developments from the speculative future. Narrow artificial intelligence (ANI), also known as weak artificial intelligence, includes systems designed to perform specific, well-defined tasks, from chess programs to image classifiers to language translators. These systems, despite sometimes superhuman performance in their fields, "lack general cognitive abilities" and cannot transfer their competencies to other problem areas (Russell and Norvig, 2021, p. 34). Every currently used artificial intelligence application, regardless of its sophistication, belongs to this category.

Artificial general intelligence (AGI), on the other hand, refers to hypothetical systems that would match or exceed human cognitive abilities in all domains. Also known as strong artificial intelligence or human-level artificial intelligence, AGI represents a long-standing quest to create machines with true understanding and flexible intelligence comparable to human cognition. The terminology itself reveals conceptual ambiguity: some researchers reserve the term “strong artificial intelligence” exclusively for systems that would possess consciousness or phenomenological experience, while others use it more broadly to indicate human-level performance, regardless of internal states. This terminological confusion reflects deeper philosophical disputes about the nature of intelligence and consciousness.

The distinction between narrow and general artificial intelligence has profound implications for how we assess the progress and risks associated with artificial intelligence. If current systems are essentially narrow, their sometimes “super-human” capabilities in limited, specific domains don’t seem to provide clear paths to achieving AGI. On the other hand, if narrow AI systems can be scaled or combined to achieve generality (as some researchers currently claim with regard to large language models) then the boundary between narrow and general AI may be more fluid than traditionally assumed. The development and potential achievement of AGI remain a subject of intense debate in the AI community, with recent advances leading some researchers to argue that early forms of AGI may already exist. It’s worth noting that naturally depends on the very definition of AGI, and none that would be widely accepted doesn’t exist. Some researchers further argue that AGI should be understood not as a threshold to be crossed but rather as a continuum.

1.1.5. Types of AI According to Technical Approach

In the history of artificial intelligence development, it may be observed how different technical approaches reflect different beliefs about what intelligence really is. In the early days—from the 1950s through the 1980s—symbolic AI (sometimes called "Good Old-Fashioned AI" or GOFAI) was the dominant approach. Researchers believed they could capture intelligence by having computers manipulate symbols according to strict rules. The thinking was simple: if human brains work by processing symbols, then intelligence is basically just computation.

But this symbolic approach started reveal its limits. It struggled with what researchers called “knowledge acquisition bottlenecks”. Basically, it was incredibly hard to feed these systems all the knowledge they needed. And when faced with messy, real-world problems, these systems often were failing. These failures made people wonder whether one could

really capture the full complexity of intelligent behavior just by writing out explicit rules and representations. Machine learning (ML) represents a paradigm shift from explicit programming to systems that "improve performance through experience" by extracting patterns from data (Mitchell, 1997, p. 2). This approach departs from the assumption that intelligence requires explicit symbolic reasoning, treating it instead as resulting from statistical patterns. Deep learning, a subset of ML that uses multilayer neural networks, has achieved remarkable success by learning hierarchical representations without explicit feature engineering. The philosophical implications are profound: if intelligence can arise from simple computational units organized in layers, what does this suggest about the nature of cognition itself?

The recent emergence of generative artificial intelligence (GenAI), exemplified by large language models such as GPT-4 and image generators such as DALL-E, represents another conceptual shift. These systems generate novel content by learning the statistical structure of vast datasets. While the results are impressive the question remains, what actually constitutes true novelty, and whether some recombination of the data that an algorithm has "seen", indeed qualifies as such. Hybrid approaches, combining neural and symbolic methods, suggest that neither of these paradigms fully captures the spectrum of intelligent behavior. Each technical approach therefore reflects a different theoretical assumption about what intelligence fundamentally is: rule-following, pattern recognition, next token prediction or something else entirely.

1.1.6. Turing's Fundamental Contribution

Alan Turing's 1950 article "Computing Machinery and Intelligence" established a revolutionary approach to defining machine intelligence that continues to shape contemporary debates. Instead of attempting to answer the seemingly unsolvable question "Can machines think?", which Turing considered "too meaningless to be worth discussing", he proposed an operational test based on observable behavior (Turing, 1950, p. 442). The Turing test, later known as the imitation game, involves testing whether a machine can mimic a human well enough in a text conversation to fool the questioner. This behaviorist reformulation shifted the focus of the definition from internal states or consciousness to external action, establishing a pragmatic criterion for intelligence that bypassed metaphysical controversies about the mind and consciousness.

Turing's approach embodied a philosophical position with far-reaching consequences: intelligence should be defined by what it does, not by what it is. This functionalist

perspective suggests that any system exhibiting intelligent behavior should be considered intelligent, regardless of its physical substrate or internal mechanisms. The Turing test operationalizes intelligence as a social and linguistic phenomenon, meaning that intelligence manifests itself through effective participation in human communication practices. This approach simultaneously democratizes intelligence (any system can qualify) while maintaining an anthropocentric standard (success is measured against human performance). The influence of Turing's behavioral criterion extends far beyond its original formulation. Contemporary comparative tests of artificial intelligence, from question-answering tasks to game competitions, follow Turing's example by defining intelligence through measurable outcomes rather than internal architecture. However, critics argue that the Turing test confuses intelligence with its appearance, potentially rewarding sophisticated mimicry while overlooking true understanding. The ongoing debate about whether large language models truly understand language or merely simulate understanding is a modern manifestation of the questions that Turing's approach sought to sidestep.

1.1.7. AGI Levels According to DeepMind

The framework proposed by Google DeepMind researchers in 2023 represents a significant evolution in how the AI community conceptualizes and measures progress toward artificial general intelligence. Morris et al. (2023) introduced levels of AGI performance, generality, and autonomy, providing a common language for comparing models and measuring progress toward AGI. This moves the debate beyond binary questions of whether a system "is" or "is not" AGI, offering instead a gradual scale that accounts for the likely uneven development of AI capabilities across different domains. This framework delineates five (or six if to include zero-level) ascending levels of capability, each defined by performance relative to fundamental human values: (0) No AGI; (1) Emerging AGI (equal to or slightly better than an unskilled human); (2) Competent AGI (at least the 50th percentile of skilled adults); (3) Expert AGI (90th percentile); (4) Virtuoso AGI (99th percentile); and (5) Superhuman AGI (exceeds 100% of humans).

Performance (rows) x **Generality** (columns)	**Narrow** *clearly scoped task or set of tasks*	**General** *wide range of non-physical tasks, including metacognitive tasks like learning new skills*
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging *equal to or somewhat better than an unskilled human*	**Emerging Narrow AI** GOFAI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	**Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai & Hassabis, 2023)
Level 2: Competent *at least 50th percentile of skilled adults*	**Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert *at least 90th percentile of skilled adults*	**Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	**Expert AGI** not yet achieved
Level 4: Virtuoso *at least 99th percentile of skilled adults*	**Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016; 2017)	**Virtuoso AGI** not yet achieved
Level 5: Superhuman *outperforms 100% of humans*	**Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	**Artificial Superintelligence (ASI)** not yet achieved

(*Figure 1*. Moris et. Al, 2023)

The DeepMind's researchers note that levels higher than emerging AGI have not yet been achieved, with current systems such as GPT-4 are classified as emerging AGI. This classification introduced more systematic approach to the debate related to defining AGI.

The philosophical significance of this framework extends beyond mere taxonomy. By explicitly considering both the performance and generality dimensions, it recognizes that AGI is not a monolithic achievement, but rather a complex phenomenon that may emerge gradually and unevenly. The framework also introduces a dimension of autonomy, recognizing that the same level of capability can manifest itself in different paradigms of interaction: from AI as a tool to AI as an autonomous agent. This multidimensional approach understands the intelligence as composed of many potentially separate capabilities, rather than as a single, uniform phenomenon. On the other hand, it represents also a great example of the AGI term ambiguity, which in has been multiple times redefined, and can mean many things depending on the context. It also raises questions about some underlying criteria for measurement. If one would like to verify whether AI achieved for instance "Competent" level, it may be challenging to prove that indeed it reached the threshold "at least 50th percentile of skilled adults". The paper also doesn't provide a clear criterion for what it

means by “skilled” adults. Therefore, although interesting, the study doesn’t provide a definitive criteria for potential AGI evaluation.

1.1.8. From Capability to Transformation

Contemporary industry leaders have shifted definitions of artificial intelligence from technical specifications toward transformative potential, emphasizing not what artificial intelligence is, but what it can become and what it can achieve, and sometimes even radically changing the AI term understanding. Sam Altman, the CEO of OpenAI in the early 2024 during the interview with MIT professor and podcaster Lex Fridman, said that to him, asked about AGI, that to him it’s furthermost about its potential to transform the reality, especially in the scientific and economic context (Fridman, 2024). Short after, also the CEO of Anthropic Dario Amodei, was speaking along the same lines. This definition reconceptualizes intelligence through the lens of productive capacity, suggesting that the measure of AI’s achievements is not cognitive equivalence but economic impact. Altman’s vision goes further, presenting AGI as an engine of radical abundance that could fundamentally restructure economic systems, potentially rendering traditional notions of scarcity obsolete. This framing transforms AGI from a scientific achievement into an economic revolution.

Early 2024, Satya Nadella, the CEO of Microsoft, set a new metric for AGI’s arrival: a 10% GDP growth rate in the developed world (Team, 2025). Moreover, as reported by media, Microsoft and OpenAI signed an agreement stating OpenAI has only achieved AGI when it develops AI systems that can generate at least $100 billion in profits (Zeff, 2024). This shifts far from the technological and philosophical definitions of AI.

1.1.9. Broader Socio-Technical Definitions

Critical scholars have developed definitions of artificial intelligence that emphasize its political and economic dimensions, challenging purely technical characteristics. Kate Crawford’s book “Atlas of AI” (2021) presents artificial intelligence not as an abstract computational capacity, but as an extractive industry embedded in specific material and social relations. Crawford argues that artificial intelligence should be understood as a manifestation of power involving the extraction of natural resources for hardware, human labor for data annotation and content moderation, and vast amounts of data derived from human activity. This definition shifts the focus from algorithms and intelligence to

infrastructure and power relations, revealing that artificial intelligence is inseparably linked to the conditions of its production.

This perspective brings to the attention the dimensions hidden by technical definitions. The focus on extraction highlights how AI systems depend on often invisible human labor, from workers labeling training data to content moderators protecting users from harmful outcomes. Crawford's work shows that AI is part of what Shoshana Zuboff calls "surveillance capitalism" (2019) in which human behavior becomes the raw material for predictive products. By looking at AI through the lens of political economy instead of just its technical features, this approach exposes the power imbalances and exploitation that make AI development possible. The extraction framework also highlights the environmental costs of AI, from rare earth minerals in computer hardware to the enormous energy consumption involved in training and running large models.

Treating AI as a tool of power, not a neutral technology, reshapes the approach to evaluating and governing it. If AI centrally concerns the distribution of power and social control, then questions of bias, fairness, and equity move from the margins to the core of development. This perspective challenges narratives of inevitable technological progress, framing AI's trajectory as a series of political choices about how society is organized and whose interests it serves. This shift matters for governance too. AI governance stops being just a technical problem that only experts can solve. It becomes a political issue that requires democratic input and public participation.

1.1.10. Definition in the EU Artificial Intelligence Act

The European Union's Artificial Intelligence Act, which came into force on August 1, 2024, is a comprehensive legal framework regulating artificial intelligence, defining an artificial intelligence system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (*Regulation - EU - 2024/1689 - EN - EUR-LEX*, n.d.). This regulatory definition proposes a specific view for understanding AI and shaping its development. By creating legal boundaries and rules, it's having a influence on how AI is being developed and used in one of the biggest markets in the world. The law focuses on three main features: machine-based operation, adaptability, and output generation. This creates a definition broad enough to cover most current AI applications while staying flexible for whatever comes next.

Additionally, the regulation groups AI systems according to how they might affect society, not based on their technical specifications. It introduces risk-based approach with four risk categories: unacceptable risk (completely banned, like social scoring systems), high risk (critical applications that need thorough assessment), limited risk (must meet transparency requirements), and minimal risk (most applications, with no special rules). The Act also requires human oversight of AI systems to keep them from causing harm. This shows a strong belief that people should remain in charge, even as we rely more and more on autonomous systems. The impact of this regulatory definition reaches beyond just legal compliance: it advocates looking at AI as a sociotechnical phenomenon instead of merely technical one.

1.1.11. Critical Perspectives: "Stochastic Parrots"

With the rise of large langue models, debates about AI definitions intensified. To many the LLMs capability of communicating with great efficiency in natural language is a clear exhibit of intelligence. These views met with a critique. In 2021, Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell published an article titled "On the Dangers of Stochastic Parrots", introducing a metaphor that depicts large language models as systems that statistically mimic text without true understanding. The paper's main argument challenges equating linguistic competence with true understanding: LLM models, despite generating coherent and contextually appropriate text, merely "probabilistically string together words and sentences without regard to meaning" (Bender et al. 2021, p. 615). The metaphor of a stochastic parrot captures this discrepancy: just as a parrot repeat sounds without understanding their meaning, LLM models reproduce linguistic patterns without capturing semantic content. This criticism goes beyond technical limitations and questions fundamental assumptions about what these systems achieve.

The article points to several risks associated with the current trajectory of LLM development. First, the environmental costs are significant: training large models requires enormous computational resources that generate a significant carbon footprint, raising questions about sustainability and environmental justice. Second, these models amplify biases present in training data, potentially perpetuating and legitimizing discrimination on a large scale. Third, the illusion of understanding created by fluent text generation can lead to misplaced trust and deployment in sensitive contexts. As the authors argue, because LLM models "simply generate outputs based on training data," they "do not understand whether they are saying something incorrect or inappropriate". The criticism highlights

multidimensional – technical, ethical, and epistemological – nature of AI, questioning not only what LLM models can do, but also what we should do with them.

1.2. Anthropomorphization of Artificial Intelligence

The rapid development of recent AI systems, especially their capability to generate human-like content, has sparked a philosophical debate about the human tendency to attribute anthropomorphic characteristics to machines. This phenomenon, though deeply rooted in the very intentions behind AI disciple origin, takes on unprecedented significance as artificial intelligence systems display increasingly sophisticated behavior that superficially resemble human cognition and agency.

Floridi and Nobre take up the topic in their paper "Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences" (2024). The authors begin their analysis by identifying a phenomenon they call "conceptual borrowings," whereby new disciplines adopt terminology from "neighboring", established fields in order to construct their technical vocabulary. This proves necessary, especially when a scientific disciplines like AI, evolve faster than they can develop their own appropriate terminology. The authors note that this linguistic migration is analogous to Carl Schmitt's observation regarding the theological origins of contemporary political concepts, suggesting a recurring pattern in the evolution of scientific disciplinary languages. In the specific context of artificial intelligence and cognitive science, these borrowings have led to a phenomenon that the authors describe as bidirectional contamination: artificial intelligence has adopted anthropomorphic descriptions of computational processes, while cognitive science increasingly uses computational metaphors to refer to biological cognition.

The consequences extend far beyond semantic imprecision. When artificial intelligence researchers describe machine learning algorithms as exhibiting "attention" or "hallucinations", they are referring to terms whose psychological meanings bear little resemblance to their computational implementations (Floridi and Nobre, 2024). For example, the "attention mechanism" of transformer architecture involves mathematical operations on vector representations that bear only distant similarities to the complex neurobiological and phenomenological processes underlying human attention.

The bidirectional dimension of this conceptual entanglement proves equally problematic. As cognitivists increasingly apply computational frameworks to explain human

cognition, they risk reducing the rich complexity of biological intelligence to algorithmic processes. Floridi and Nobre argue that this reductionist tendency impoverishes our understanding of uniquely human capacities. The authors point to three groups particularly affected by this confusion: those who sincerely believe that current artificial intelligence possesses intelligence, those who predict the emergence of superintelligence in the future, and those who exploit conceptual ambiguity for commercial or ideological purposes.

While Floridi and Nobre notice challenges, Gunkel rather than viewing anthropomorphism as a cognitive error requiring correction, argues that robots occupy a unique ontological position that disrupts the traditional person/thing dichotomy underlying Western moral and legal thought (Gunkel, 2023). His analysis suggests that attempts to classify robots as quasi-persons deserving of rights or mere objects subject to unlimited manipulation fail to capture the distinctive nature of the human-robot relationship. Gunkel's approach draws on Emmanuel Levinas' philosophy of otherness, suggesting that moral considerations can arise from encounters with entities that present themselves as social others, regardless of whether they possess consciousness or other internal properties. This relational structure sidesteps difficult questions about machine consciousness, focusing instead on the phenomenology of human-robot interaction. Gunkel argues that when humans encounter robots that exhibit social presence, moral obligations may arise from the structure of that encounter rather than from the properties of the robot itself (Gunkel, 2023).

The implications of Gunkel's relational approach extend beyond theoretical philosophy to practical questions about the rights of robots and their legal status. By arguing that moral considerations depend on social relations rather than ontological properties, Gunkel provides a framework for addressing emerging ethical challenges without having to resolve metaphysical questions about machine consciousness. This perspective is particularly relevant in the context of the growing participation of social robots in roles traditionally reserved for humans in healthcare, education, and companionship.

Critics of Gunkel's position, notably Bryson (2018) and Sætra (2021), argue that his relational approach risks equating appearance with reality, which could lead to the misallocation of moral concern and resources. They argue that granting moral status based on superficial social signals rather than genuine consciousness may diminish the importance of consciousness and suffering. Nevertheless, Gunkel's work has had a significant impact on contemporary debates, showing that traditional moral frameworks may be inadequate for addressing the new challenges posed by advanced artificial agents.

Another point of view is presented by Coeckelbergh's phenomenological approach to human-robot interaction. It provides key insights into how anthropomorphism shapes relationships of trust with artificial systems. His analysis starts from the observation that robots phenomenologically appear to be more than mere machines, presenting themselves as social entities despite their artificial nature (Coeckelbergh, 2012). He argues that this appearance has ethical significance independent of whether robots actually possess consciousness or agency. Coeckelbergh's concept of "quasi-trust" offers a sophisticated framework for understanding how humans navigate relationships with artificial agents. Unlike trust in humans, which typically assumes intentionality and moral responsibility, quasi-trust in robots requires what the author calls "dual awareness": recognition of the artificial nature of the robot combined with a willingness to engage in trust-based interactions (Coeckelbergh, 2012). This double consciousness enables beneficial cooperation between humans and robots, while maintaining appropriate epistemic humility regarding the actual capabilities of robots.

The phenomenological perspective reveals that anthropomorphism operates not only at the level of conscious beliefs, but also through pre-reflective engagement with the world. When humans interact with social robots, they respond to social cues through embodied habits and expectations developed through human-human interactions. These reactions occur before reflective assessment of the nature of the robot, suggesting that anthropomorphism cannot be eliminated solely through rational analysis. Coeckelbergh's work challenges approaches that treat anthropomorphism as a mere cognitive error that can be corrected through education.

Furthermore, Coeckelbergh (2021) argues that different cultural contexts shape anthropomorphic responses in different ways. Eastern philosophical traditions, with less rigid boundaries between animate and inanimate beings, may foster different patterns of human-robot interaction than Western frameworks based on Cartesian dualism. This cultural dimension complicates universal recommendations for appropriate responses to artificial agents and suggests the need for a culturally sensitive approach to robot design and deployment.

Joanna Bryson's thesis that "robots should be slaves" represents the most uncompromising rejection of anthropomorphic frameworks in contemporary AI philosophy. Her position is based on a fundamental distinction between beings that are born and those that are made: robots, which are designed and manufactured, belong entirely to their creators and users in a way that biological beings do not (Bryson, 2010). She argues that this

relationship of ownership precludes the possibility of artificial agents obtaining true moral status.

Bryson's argument goes beyond the issue of ownership and encompasses the ethical implications of creating beings with apparent moral status. She argues that constructing robots to elicit anthropomorphic responses is a form of deception that undermines human dignity and authentic relationships (Bryson 2018). When people form emotional bonds with beings incapable of reciprocating care, they diminish their capacity for genuine human relationships. This problem is particularly important in contexts where robots can replace human caregivers or companions. The practical implications of Bryson's position include strong opposition to granting legal personhood to artificial intelligence systems and skepticism about the applications of social robotics. She argues that attributing responsibility or rights to robots allows humans to avoid responsibility for the consequences of automated systems. Anthropomorphic approaches that suggest otherwise serve to obscure human responsibility behind a veil of artificial agency (Bryson 2018).

Critics of Bryson's position argue that her clear distinction between humans and robots may become less and less valid as artificial intelligence systems develop. Nevertheless, her work provides important counterarguments to free anthropomorphism and raises key questions about the ethical implications of creating entities that blur traditional ontological boundaries. Her emphasis on maintaining a clear distinction between humans and machines is an important corrective to the uncritical acceptance of anthropomorphic approaches.

The question of the moral status of artificial entities is perhaps one of the most controversial aspects of the anthropomorphism debate. The spectrum of philosophical positions ranges from outright denial of moral treatment of machines to arguments for potential rights for robots based on social relationships or functional capabilities. This diversity reflects deeper disputes about the basis of moral status: whether it derives from internal properties such as consciousness, from interpersonal relationships, or from social conventions.

1.2.1. The epistemological challenge

The epistemological dimensions of the anthropomorphization of artificial intelligence pose profound challenges to both scientific understanding and practical decision-making. Floridi and Nobre have identified conceptual linkages that reveal how linguistic ambiguities cause systematic misunderstandings about the nature of intelligence, consciousness, and agency.

These misunderstandings extend beyond academic discourse and influence public perception, policy decisions, and resource allocation in ways that could profoundly affect the technological future of humanity.

The challenge of determining what artificial intelligence systems actually are, as opposed to how they present themselves, proves particularly troublesome given the opacity of contemporary machine learning systems. Deep neural networks operate on processes that are obscure to human interpretation, making it difficult to assess whether seemingly intelligent behavior reflects true understanding or is the result of sophisticated pattern matching. This epistemological limitation complicates efforts to establish clear boundaries between anthropomorphic projection and accurate recognition of emerging capabilities. Furthermore, the rapid pace of AI development creates moving targets for philosophical analysis. Abilities once considered uniquely human, such as creative expression or strategic reasoning, now *appear* to be within the reach of artificial systems. This technological dynamic challenges static philosophical frameworks and requires an adaptive approach that can account for actual progress while maintaining a critical eye on anthropomorphic projections. The epistemological challenge, therefore, lies not only in understanding current artificial intelligence, but also in anticipating future changes without succumbing to unfounded optimism or excessive skepticism.

Design decisions regarding the degree and type of anthropomorphic features in AI systems require careful ethical analysis. While some applications, such as therapeutic robots for dementia patients, may benefit from controlled anthropomorphic design, other contexts require transparent representation of artificial nature. Designers must navigate between leveraging the beneficial aspects of anthropomorphic response and avoiding deception or manipulation. Navigating this space requires interdisciplinary collaboration between technologists, ethicists, psychologists, and domain experts.

1.3. Ethics of Artificial Intelligence

The emergence of artificial intelligence as a transformative force in contemporary society has created an urgent need for reflection and guidelines on ethics. In response the whole discipline of Artificial Intelligence ethics emerged as way of systematic reflection and guidelines for shaping AI design and development. This nascent discipline operates at the intersection of technological innovation and moral philosophy, attempting to navigate the complex terrain between computational capabilities and human values. Coeckelbergh

emphasizes that ethical reflection on artificial intelligence must go "beyond mere media hype and nightmare scenarios to address concrete questions" about implementation, governance, and social impact (Coeckelbergh 2020, p. 15). This pragmatic orientation distinguishes contemporary AI ethics from speculative discussions about artificial general intelligence or science fiction scenarios. Instead, the field focuses on immediate challenges: algorithmic biases affecting marginalized communities, privacy violations through ubiquitous surveillance, the opacity of automated decision-making systems, and the environmental costs of computing infrastructure.

At the heart of this discipline are several fundamental questions that shape ongoing debates and research. How can we ensure that AI systems respect human dignity and fundamental rights while maximizing their beneficial applications? What constitutes fairness and justice when decisions are delegated to algorithmic processes? How can we maintain meaningful human agency and oversight in an increasingly automated world? Who is responsible when AI systems cause harm or perpetuate injustice? Addressing those questions require ongoing interdisciplinary dialogue between philosophers, computer scientists, policymakers, and the communities affected.

1.3.1. Key Challenges in AI Ethics

Bias and Discrimination: Perpetuating Social Inequalities

The challenge of bias in artificial intelligence systems is one of the most pressing ethical issues in this field. As Floridi and Taddeo (2016) state, opacity and bias are key issues in what is now sometimes called 'data ethics' or 'big data ethics'. This bias manifests itself in many ways: training data that reflects historical prejudices, algorithms that unintentionally discriminate, and implementation contexts that exacerbate existing inequalities. The consequences of this phenomenon are far from abstract and affect real people's access to employment, credit, healthcare, and justice. When AI systems trained on biased data make decisions about parole, loan applications, or medical treatment, there is a risk that, under the guise of computational objectivity, they will perpetuate systemic discrimination.

The sources of bias in AI systems are multidimensional and often interdependent. Training data bias occurs when datasets reflect historical patterns of discrimination or fail to adequately represent marginalized groups. For example, facial recognition systems trained primarily on images of white men show a significantly higher error rate when identifying women and people of color (Gebru et al. 2018). Algorithmic bias derives from mathematical

formulas and optimization criteria built into machine learning models that may prioritize certain outcomes or populations over others. Representational bias occurs when entire groups are absent or underrepresented in data sets, leading to poor or unfair performance of systems with respect to those populations.

The ethical implications of algorithmic bias extend beyond individual harm and encompass broader issues of social justice and democratic participation. When AI systems systematically discriminate against certain groups, they contribute to what Crawford (2021) calls "the reinforcement of historical inequalities through computational means". This reinforcement effect is particularly troubling given the scale and speed of AI systems, which can potentially affect millions of people before bias is detected and removed. Furthermore, the perceived objectivity of algorithmic decisions can legitimize discriminatory outcomes, making them more difficult to challenge than human biases.

Addressing bias in AI requires more than just technical fixes; it requires a fundamental rethinking of how we perceive fairness and justice in a computational context. Coeckelbergh (2020, p. 78) argues that the ethical discussion can proceed along the lines of formulating principles that are meant to provide guidance on what to do from a moral standpoint. However, translating abstract principles of fairness into concrete algorithmic implementations proves difficult, as different mathematical definitions of fairness often contradict each other. This tension highlights the inherently value-laden nature of AI development and the impossibility of finding purely technical solutions to ethical problems.

Privacy and Surveillance: The Erosion of Personal Autonomy

The proliferation of AI-based surveillance technologies poses an unprecedented challenge to privacy and personal autonomy. Contemporary AI systems enable forms of monitoring and analysis that go beyond traditional notions of surveillance, creating what Zuboff (2019) calls "surveillance capitalism": an economic system based on the collection and analysis of data about human behavior. These systems aggregate various data sources to create detailed profiles of individuals, predict future behavior, and influence decision-making in ways that fundamentally undermine the concepts of privacy, consent, and human agency. The ethical implications go beyond violations of individual privacy and include broader concerns about social control and democratic freedom.

The use of AI in state surveillance raises particularly serious ethical concerns about the balance between security and freedom. Facial recognition systems used in public spaces enable unprecedented tracking of individuals' movements and connections, which can

restrict freedom of expression and assembly. In authoritarian contexts, these technologies facilitate the targeting of dissidents and minorities, as documented by the use of artificial intelligence in China for ethnic profiling and social control. Even in democratic societies, the use of predictive policing algorithms and automated surveillance systems raises concerns about the presumption of innocence, fair trial, and the right to privacy.

Transparency and Explainability: The Black Box Problem

The opacity of many AI systems, particularly deep learning models, poses a fundamental challenge to accountability, trust, and human agency. Whittaker et al. (2018, p. 18) note that artificial intelligence systems used for automated decision support and 'predictive analytics' raise serious concerns about lack of due process, accountability, community involvement, and auditability. The "black box" problem extends beyond technical opacity to encompass the broader ecosystem of AI development and deployment, where proprietary algorithms, complex supply chains, and diffuse accountability obscure decision-making mechanisms. The consequences are particularly serious in high-stakes areas such as criminal justice, healthcare, and financial services.

The requirement for transparency in AI systems reflects many ethical imperatives. From a deontological perspective, individuals have a right to understand decisions that significantly impact their lives, especially when those decisions are made by or with the aid of automated systems. From a consequentialist perspective, transparency enables the identification and correction of errors, biases, and unintended consequences. From a virtue ethics perspective, transparency cultivates the institutional virtues of honesty, accountability, and credibility. However, achieving meaningful transparency proves difficult in technical and conceptual terms, as the mathematical operations underlying neural networks are resistant to intuitive human interpretation.

The tension between explainability and performance is a fundamental trade-off in AI development. More interpretable models, such as decision trees or linear regression, often achieve lower accuracy than opaque deep learning systems. This trade-off forces difficult choices between competing values: should we prioritize systems that we can understand, or systems that perform better according to narrow metrics? The answer depends on context, stakes, and values. Recognizing that explainability is not merely a technical property, but an ethical and political choice about the kinds of systems we choose to deploy and the forms of accountability we demand.

Environmental Impact: The Hidden Costs of Computation

The environmental impacts of AI development and deployment are an increasingly urgent ethical concern that remains under-explored in mainstream discourse. Training large-scale AI models requires enormous computational resources, generating carbon emissions comparable to those of small cities. (Strubell et al., 2019) calculated that training a single large language model can emit as much carbon dioxide as five cars over their entire lifetime. The environmental cost must be considered in the ethical assessment of AI systems, especially given the climate crisis and the need for sustainable technological development. The carbon footprint of AI extends beyond training to include inference, data storage, and the production of specialized hardware.

The geographic and social distribution of AI's environmental impact raises questions of environmental justice. Data centers are often located in regions where electricity is cheap, often derived from fossil fuels, while the benefits of AI services are primarily enjoyed by wealthy countries and individuals. Crawford and Joler (2018) trace the "anatomy of an AI system," revealing complex supply chains involving rare earth mining, manufacturing, and e-waste disposal that disproportionately impact communities in the global South. These hidden costs complicate the narrative of AI as a clean, dematerialized technology and underscore the need to assess the life cycle of AI systems.

The relationship between AI and sustainability presents both a challenge and an opportunity. While AI's energy consumption contributes to climate change, its applications in energy optimization, climate modeling, and environmental monitoring offer potential benefits. This dual nature requires careful ethical analysis to distinguish genuine environmental applications from "greenwashing," which uses sustainability rhetoric to justify further expansion of computing infrastructure. The principle of proportionality suggests that the environmental costs of AI should be weighed against its benefits.

Work and Employment: The Human Costs of Automation

AI-based systems threaten to replace work in many sectors, from manufacturing and transportation to professional services such as law and medicine. With the invention of Generative AI architectures also all the work done in front of a computer is arguably at the risk of significant automation. The distributional effects of AI-based automation exacerbate existing inequalities. Furthermore, the benefits of automation, namely increased productivity

and profits, flow primarily to capital owners rather than workers, raising fundamental questions about the distribution of value generated by AI systems.

In addition to fear of job losses, AI is changing the very nature of work, often in ways that raise ethical concerns. Algorithmic management systems subject workers to constant surveillance, automatic performance evaluation, and opaque disciplinary procedures. Platform workers in the gig economy face precarious employment conditions, guided by algorithms that optimize productivity rather than worker well-being. Even in traditional employment settings, artificial intelligence systems that monitor performance, predict employee behavior, or automate hiring and promotion decisions raise concerns about dignity, autonomy, and fairness in the workplace. On the other hand, as some argue, disruption of work does not necessarily mean complete job replacement; rather, it means the transformation of work itself. AI has the potential to empower workers to focus on more creative tasks and, in economic terms, to enable what some call “one person unicorns”, meaning single entrepreneurs valued at $1 billion (Przegalińska & Triantolo, 2024, p. 70). Even if that promise holds, the beneficiaries of the transformation are most likely those who are already better educated and wealthier. An ethical response to AI’s impact on work will require rethinking social contracts and economic systems.

Other Societal Risks

Artificial intelligence also brings other significant risks to the social environment, including the flooding of media and communication channels with deepfakes, difficulties in the control of autonomous weapons, and the empowerment of malicious actors. It’s hard to predict all the dangers, as they vary from amplifying and automating issues already existing in society, to some unknown unknowns, brought by the specific of those technologies. That said, it may be difficult to imagine a single industry or social dimension which wouldn’t be impacted by the advent of powerful AI systems.

1.3.2. Ethical frameworks

As way of addressing ethical challenges posed by the functioning and implications of AI technologies multiple institutions and organizations proposed their ethical frameworks. They consist of set of ethical guidelines to govern the design and implementation AI systems. The AlgorithmWatch website, which catalogued such guidelines, as of April 2020, listed 167 such documents (AlgorithmWatch 2020). Two of these frameworks are presented here as examples to demonstrate approaches to framing ethical guidelines. In particular, they

highlight the sociotechnical nature of AI systems and the selection and definition of ethical values, which is relevant to the context of this dissertation.

EU's Ethics guidelines for trustworthy AI

The European Union's "Ethics Guidelines for Trustworthy AI", published by the High-Level Expert Group on Artificial Intelligence in 2019, are among the most comprehensive and influential frameworks on the ethics of artificial intelligence. According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment (High-Level Expert Group on Artificial Intelligence, 2019, p. 5).

This three-pronged basis recognizes that legal compliance alone cannot ensure trustworthiness, and that purely technical solutions do not take into account the social and ethical dimensions. The framework's holistic approach recognizes AI systems as socio-technical assemblies requiring integrated governance strategies.

The guidelines set out seven key requirements that AI systems must meet in order to achieve trustworthiness, each of which addresses specific ethical issues while contributing to the overall integrity of the system.

- **Human agency and oversight**: AI should enhance human autonomy and fundamental rights by enabling informed decisions and by ensuring appropriate oversight through human in the loop, human on the loop, and human in command arrangements.
- **Technical robustness and safety**: AI should be resilient and secure. It must operate safely with fallback plans for failures and deliver accuracy, reliability, and reproducibility to minimize unintended harm.
- **Privacy and data governance**: Respect privacy and data protection, and implement strong data governance that ensures data quality, integrity, and legitimate access.
- **Transparency:** Make data, systems, and business models understandable and traceable. Provide explanations suited to each audience, and make clear when people interact with AI and what its capabilities and limits are.

- **Diversity, non-discrimination, and fairness**: Prevent unfair bias that can marginalize groups or reinforce discrimination. Ensure accessibility and include relevant stakeholders throughout the system life cycle.
- **Societal and environmental well-being**: Aim for benefits that extend to people now and in the future. Promote sustainability, consider impacts on the environment and other living beings, and weigh broader social effects.
- **Accountability**: Put mechanisms in place that assign responsibility for AI systems and outcomes. Enable audits of algorithms, data, and design processes, especially in critical uses, and provide adequate and accessible routes for redress.

These requirements form an interconnected framework in which each element reinforces the others (High-Level Expert Group on Artificial Intelligence 2019, p.14).

The philosophical basis of the EU guidelines is rooted in the fundamental rights signed into the EU treaties and the Charter of Fundamental Rights. This rights-based approach assumes that ethical requirements derive from legally recognized principles, while going beyond minimum legal requirements to encompass broader ethical aspirations. The guidelines explicitly link each requirement to relevant fundamental rights, showing how design choices impact human dignity, freedom, equality, and justice. Relying on fundamental rights provides normative authority while facilitating integration with existing legal frameworks. Critics argue, however, that a rights-based approach may reflect specific European values that may not translate to other cultural contexts.

The implementation of the EU guidelines through the Assessment List for Trustworthy Artificial Intelligence (ALTAI) is a significant attempt to bridge the gap between abstract principles and practical implementation. Published in 2020 after extensive pilot testing and stakeholder feedback, ALTAI is a detailed self-assessment tool that translates each requirement into specific questions and comments. The tool helps organizations systematically assess their AI systems, identifying potential ethical risks and suggesting mitigation strategies. This structured approach addresses a key criticism of principle-based frameworks: the lack of concrete guidance for practitioners.

The ALTAI project reflects the lessons learned from a pilot process in which various organizations from different sectors tested the assessment framework. Feedback revealed tensions between comprehensiveness and usability, with organizations finding it difficult to strike a balance between thorough assessment and resource constraints. The final version attempts to accommodate different organizational contexts and AI applications while maintaining consistent assessment criteria. The flexibility of the tool allows it to be adapted

to specific use cases, although this flexibility potentially enables selective compliance, with organizations emphasizing convenient requirements while neglecting more difficult ones.
The assessment process covers both technical and organizational dimensions, recognizing that trustworthy AI requires appropriate governance structures that go beyond technical features. Questions address not only algorithmic properties, but also development processes, organizational policies, and stakeholder engagement practices. This comprehensive scope reflects the understanding that ethical failures often stem from organizational factors rather than purely technical ones. The assessment examines how decisions are made, who participates in development, and what accountability mechanisms exist. This focus on organizational aspects distinguishes ALTAI from purely technical assessment tools, while emphasizing the importance of institutional context.

The Rome Call for AI Ethics

The Rome Call for AI Ethics represents a unique intervention in AI governance, resulting from an unexpected alliance between religious institutions and technology companies. Initiated by the Pontifical Academy for Life and first signed in February 2020, the document brings together the Vatican, major technology corporations, including Microsoft and IBM, and government bodies in a shared commitment to the ethical development of artificial intelligence (Paglia, 2024). Father Paolo Benanti, professor of a moral theology and a Franciscan with an engineering background, played one of key roles in developing and promoting this framework. His dual expertise in theology and technology enabled a distinctive synthesis that grounds technical considerations in deeper questions about human dignity and purpose.

The Rome Call sets out six core principles that signatories commit to uphold: transparency, inclusiveness, accountability, impartiality, reliability, and privacy-respecting security. While these principles overlap with other frameworks, their interpretation through the lens of human dignity and the common good gives them particular emphasis. For example, the principle of inclusion goes beyond non-discrimination and encompasses a positive vision that “the needs of all people should be taken into account so that everyone can benefit and all are provided with the best conditions for self-expression and development” (Rome Call 2020, 3). This wording reflects the emphasis in Catholic social teaching on the integral development of the human person and the preferential option for the marginalized.

The brevity of the document (three and a half pages) contrasts sharply with more elaborate frameworks, but this brevity reflects strategic choices about accessibility and cross-cultural communication. Rather than detailed technical specifications, the Rome Call offers what Benanti describes as "an offer of value for individuals, companies, and society" (Deign, 2024). This approach prioritizes establishing common ethical ground among diverse stakeholders over prescriptive requirements. The framework functions more as a covenant or commitment than a regulatory instrument, relying on moral authority and public accountability rather than law enforcement.

The theological and philosophical foundations of the Rome Call draw on the tradition of natural law, which affirms universal human dignity and common moral principles accessible through reason. This universalistic approach enables dialogue across religious and secular divides, as evidenced by the expansion of the framework to include Jewish and Muslim leaders in 2023. The emphasis on human dignity as a fundamental principle from which other requirements derive is a unifying concept that transcends specific cultural or religious contexts. However, this universalism must grapple with the tension between affirming shared values and respecting legitimate pluralism in their interpretation and application.

The evolution of the Rome Call from a Catholic initiative to an interfaith and multi-stakeholder platform represents a significant advance in AI governance. The signing in January 2023 by representatives of Judaism and Islam, including the Chief Rabbi of Israel and the Abu Dhabi Peace Forum, established what the participants called an "Abrahamic commitment" to AI ethics. A meeting in Hiroshima in July 2024 expanded participation to eleven world religions, demonstrating the framework's ability to facilitate dialogue across different spiritual traditions. This religious convergence around AI ethics suggests that technological challenges can catalyze unexpected alliances and shared moral goals.

Incorporating religious perspectives into debates on AI ethics brings a distinctive contribution that is often absent in secular frameworks. Religious traditions offer rich resources for reflection on human dignity, moral responsibility, and the common good that go beyond utilitarian calculations or rights-based approaches. The concept of stewardship, common to Abrahamic traditions, provides a framework for understanding human responsibility for technology, emphasizing care and accountability rather than mere ownership or control. The religious emphasis on transcendence and eternity offers a perspective on technological change that resists both uncritical acceptance and fearful rejection of innovation.

The participation of companies in the Rome Call reflects a growing awareness that ethical credibility requires the involvement of various stakeholders outside the technical and business communities. Companies such as Cisco, which signed the call in 2024, are aligning their AI policies with the framework of the Rome Call, recognizing that "technology must be based on a foundation of trust at the highest level to ensure an inclusive future for all" (*Press Release: Cisco Signs the Rome Call for AI Ethics*, 2024). The involvement of businesses in collaboration with religious institutions may seem out of place, but it reflects a pragmatic recognition that the sustainable development of artificial intelligence requires broad social acceptance that cannot be achieved through technical expertise alone.

The challenges of maintaining consistency while accommodating diversity become apparent during the implementation of the framework. Different signatories interpret the principles through the lens of their specific perspectives, which can lead to discrepancies in their application. For example, the principle of inclusiveness may be understood differently by technology companies focused on market access, governments concerned with the well-being of their citizens, and religious institutions emphasizing the spiritual dimension of human development. The lack of enforcement mechanisms or detailed specifications allows for such interpretive flexibility, but risks reducing the framework to a symbolic rather than substantive commitment. Effectiveness depends on ongoing dialogue and mutual accountability among signatories.

A critical analysis of the Rome Call reveals both limitations and a distinctive contribution to the discussion on AI ethics. The conciseness and generality of the framework, while enabling broad agreement, provide limited practical guidance for implementation. Terms such as "transparency," "reliability," and "impartiality" remain undefined, allowing signatories to claim compliance while continuing problematic practices. As one analysis notes, "the openness of the statement does not allow for the determination of intent" with regard to specific requirements (Aif, 2025). This ambiguity may facilitate initial agreement but complicates accountability and compliance assessment. The voluntary nature of the Rome Call raises doubts about its effectiveness in limiting harmful applications of artificial intelligence.

Despite these limitations, the Rome Call brings several distinctive elements to the discussion on AI governance. Its emphasis on human dignity as a fundamental principle provides a robust normative foundation that goes beyond purely economic or technical indicators. The religious underpinnings of the framework provide moral authority that purely secular initiatives may lack, particularly in communities where religious institutions have

significant influence. The interfaith dimension shows that diverse spiritual traditions can find common ground on technological issues, potentially facilitating global cooperation that transcends cultural and political divides.

The vision of "algorethics" contained in the Rome Call – Benanti's term for the ethical dimensions of algorithmic decision-making – is a conceptual contribution that goes beyond specific rules. This neologism captures the idea that ethics must be built into the very structure of algorithmic systems, rather than applied as an external constraint. This concept suggests that ethical considerations should influence every stage of AI development, from problem formulation to data collection, model design, implementation, and evaluation. This integrated approach challenges the separation of technical and ethical dimensions that characterizes most AI development.

1.3.3. AI ethics, trustworthy AI, and responsible AI: conceptual distinctions

The terms *AI ethics*, *responsible AI*, and trustworthy AI, while often used interchangeably, represent distinct but complementary layers in a coherent governance hierarchy. This framework moves from abstract principles to concrete practices and ultimately to verifiable outcomes, providing a structured way to navigate the complexity of AI governance.

AI ethics is a broad philosophical field concerned with the moral principles guiding the development of artificial intelligence. It asks the fundamental question: "Why is this the right thing to do?" Drawing on established ethical theories, it provides a normative compass for the entire field, establishing basic principles such as fairness, transparency, and accountability. These principles are universal and abstract, providing the moral foundation for all further action. However, their general nature can make it difficult to translate them directly into specific engineering practices, sometimes leading to criticism of "ethical washing" when they are not backed up by concrete actions.

Responsible AI (RAI) is based on this ethical foundation, providing an organizational and procedural framework for implementation. RAI answers the practical question: "How should AI be properly built and implemented?" This concept is primarily used by corporations and developers to translate abstract ethical ideals into concrete corporate policies, engineering practices, and governance structures. RAI involves implementing ethics throughout the entire AI lifecycle, from data acquisition to model deployment and monitoring. This includes establishing internal review committees, using technical tools to detect and mitigate bias, and promoting a culture of ethical awareness

within the organization. Essentially, RAI is the engine that transforms ethical principles into a repeatable, manageable process.

Trustworthy AI (TAI) represents the desired, verifiable, and systemic outcome. TAI shifts the focus from the developer's internal processes to the external, auditable properties of the AI system itself. It answers the key question: "What are the auditable characteristics of a properly constructed system?" This concept is preferred by regulatory and standardization bodies such as the European Union and the National Institute of Standards and Technology (NIST) in the United States because it defines a trustworthy system as one that is demonstrably lawful, ethical, and robust. The TAI framework provides measurable criteria and checklists for assessment, creating a basis for regulation, certification, and building public trust.

Significant similarities between these concepts comes from shared vocabulary. Principles such as fairness, accountability, and transparency are central to all three. However, the higher up the hierarchy, the more granular they become. In AI ethics, fairness is a philosophical ideal. In responsible AI, it becomes a process requirement, demanding actions such as the use of diverse training data. In trustworthy AI, fairness is a verifiable property of the system that can be measured using statistical indicators and audited against a standard.
The key differences are in their scope, audience, and outcomes. AI ethics is philosophical in nature and addresses society and policymakers, providing them with guidance. Responsible AI is procedural in nature and addresses developers and corporations, providing them with internal rules and toolkits. Trustworthy AI is results-oriented and aimed at users, regulators, and auditors, providing them with measurable standards and risk management frameworks.

1.3.4. The Challenge of Implementation: From Principles to Practice

Translating abstract ethical principles into concrete technical practices is one of the biggest challenges in AI ethics. Despite the proliferation of ethical frameworks, practitioners find it difficult to implement high-level principles such as fairness, transparency, and accountability in the actual development of systems. As Morley et al. (2021), "the theory of AI ethics remains highly abstract and has limited practical application for those actually responsible for designing AI algorithms and systems". This implementation gap stems from multiple sources: the inherent ambiguity of ethical concepts, the complexity of technical systems, the diversity of application contexts, and the lack of clear indicators of ethical compliance.
The challenge of operationalization reflects deeper tensions between the universality of ethical principles and the specificity of technical implementation. Ethical principles such as

fairness or transparency are open to multiple interpretations and may conflict when translated into technical specifications. For example, fairness can be operationalized as demographic parity, equal opportunity, or individual justice: each of which leads to different technical implementations that may be incompatible with each other. Choosing between these operationalizations involves value judgments that cannot be resolved by technical means alone. This irreducible value-laden nature of implementation decisions challenges narratives of purely technical solutions to ethical problems.

The organizational context of AI development complicates operationalization efforts. Most AI systems are the result of complex organizational processes involving multiple actors with potentially conflicting interests and values. Engineers focus on technical performance, product managers prioritize user engagement, legal teams emphasize regulatory compliance. Incorporating ethical considerations into this already complex process requires not only technical tools, but also organizational changes, new roles and responsibilities, and cultural shifts. The challenge extends beyond individual awareness to include institutional structures and incentive systems that may not reward ethical reflection.

The time dynamics of AI development create additional challenges for operationalization. AI systems evolve through iterative processes of development, testing, and refinement. Ethical issues can arise at any stage, from initial problem formulation, through data collection, model development, implementation, and ongoing operation. Furthermore, systems that appear ethical at the development stage may exhibit problematic behavior when deployed at scale or in unexpected contexts. This temporal complexity requires not a one-time ethical assessment, but continuous monitoring and adjustment. Implementation must therefore include not only initial design, but also ongoing management throughout the system's lifecycle.

Technical approaches to operationalizing AI ethics focus on developing tools, methods, and frameworks that incorporate ethical considerations into system design and operation. An example of such an approach is research on fair machine learning, which develops mathematical formulas for fairness and algorithms that optimize these metrics. Techniques such as bias elimination, fairness constraints, and demographic parity optimization aim to reduce discriminatory outcomes. Technical solutions offer concrete, measurable approaches to eliminating bias, providing developers with practical tools rather than abstract principles. However, they also reveal the limitations of purely technical approaches to ethical problems.

Explainability and interpretability techniques are another important element of technical operationalization. Methods ranging from simple feature importance scores to advanced explanatory frameworks such as LIME and SHAP aim to increase the transparency of black-box models. These techniques allow us to identify the features that have the greatest impact on predictions, visualize decision boundaries, and generate human-understandable explanations for individual predictions. However, the relationship between technical explainability and meaningful human understanding remains difficult. Explanations that meet technical criteria may still fail to provide insight that enables meaningful human oversight.

Privacy-preserving techniques offer technical mechanisms to protect the privacy of individuals while enabling the development and deployment of artificial intelligence. Differential privacy provides mathematical guarantees about information disclosure, and federated learning enables model training without centralizing sensitive data. Homomorphic encryption allows computations to be performed on encrypted data, preserving privacy throughout the analytical process. These techniques demonstrate how technical innovations can solve ethical problems, although they also involve trade-offs between privacy protection and model performance. The implementation of privacy protection techniques requires careful consideration of threat models, privacy budgets, and specific threats in particular application contexts.

Resilience and security techniques address concerns about the reliability and safety of AI systems. Adversarial training improves the resilience of models to malicious inputs, and uncertainty quantification helps identify situations where models operate outside their competence. Formal verification methods aim to provide mathematical guarantees that the system will behave within specified bounds. These technical approaches to security and resilience are necessary but not sufficient for responsible AI, as they address only certain types of risk, potentially creating a false sense of confidence in the reliability of the system. The challenge is to combine technical robustness with broader considerations of social and ethical robustness.

Process-based approaches to implementing AI ethics focus on governance mechanisms, development practices, and institutional solutions rather than purely technical solutions. Impact assessment frameworks, adapted from environmental and privacy fields, provide structured processes for identifying and assessing potential ethical risks. These assessments typically involve stakeholder consultation, risk analysis, and documentation of risk mitigation strategies. The EU's proposed requirements for high-risk AI systems are an

example of such an approach, imposing a comprehensive assessment and documentation obligation throughout the development lifecycle. However, the effectiveness of impact assessments depends on organizational commitment, expertise, and the quality of stakeholder engagement.

The establishment of AI ethics committees and review boards is an institutional approach to implementation. These bodies, modeled on institutional review boards in medical research, provide oversight and guidance for AI development projects. They typically consist of a variety of stakeholders: ethicists, domain experts, affected communities, who evaluate proposed systems for ethical issues. Large technology companies have established such committees with varying degrees of independence and authority. The effectiveness of these governance mechanisms depends on their composition, mandate, resources, and relationship to decision-making processes. Critics argue that corporate ethics committees may serve to legitimize rather than constrain the development of artificial intelligence.

The challenge is to maintain substantive ethical engagement while meeting development deadlines and performance goals. Documentation and audit trails create accountability mechanisms and enable external oversight. Model cards document the intended use, performance characteristics, and limitations of AI models.
Data sheets for datasets provide standard documentation on data collection, processing, and potential systematic errors. These documentation practices support transparency and enable end users to make informed decisions about system deployment. Audit trails store records of development decisions, creating accountability for choices made during the development process. While documentation alone cannot ensure ethical AI, it provides the necessary infrastructure for accountability and oversight.

The concept of “ethics as a service” is an emerging paradigm for scaling ethical expertise across organizations. As proposed by Morley et al. (2021) this model provides on-demand access to ethical expertise, tools, and processes, much like other enterprise services. Ethics specialists collaborate with development teams to identify risks, suggest risk mitigation strategies, and facilitate stakeholder engagement. This service model addresses the shortage of ethics expertise while avoiding the need for every developer to become an ethicist. However, there is a risk that ethics will be reduced to a compliance function rather than fostering genuine ethical reflection within organizations.

Value-sensitive design (VSD) offers a comprehensive framework for incorporating human values into the technology design process. Originating with Friedman et al. (2003),

VSD provides methods for identifying stakeholder values, understanding tensions between values, and embedding values into technical systems. The approach emphasizes iterative stakeholder engagement, conceptual exploration of values, and empirical testing of the impact of designs on values in practice. The strength of VSD lies in its systematic approach to value integration, although critics argue that it may favor certain values or stakeholders while obscuring power dynamics in the design process.

Incorporating ethical considerations into machine learning operations (MLOps) is a practical approach to operationalization. By embedding ethical checks into automated deployment processes, organizations can ensure consistent application of ethical standards. Automated bias testing, fairness monitoring, and performance tracking across different demographics become part of standard deployment procedures. This integration leverages existing DevOps practices and tools, reducing the additional burden of ethical compliance. However, automating ethical controls carries the risk of reducing complex ethical issues to simple metrics, which may result in subtle or emerging ethical issues being overlooked.

Participatory and integrative design approaches emphasize the involvement of communities affected by AI development. These methodologies, which draw on participatory action research and community-based design traditions, treat community members as partners rather than subjects in AI development. Techniques such as design workshops, community advisory boards, and joint problem definition ensure that AI systems reflect the values and needs of the community. This approach addresses concerns about imposing AI systems on communities without their participation. However, meaningful participation requires resources, time, and power sharing, which can conflict with commercial development pressures and timelines.

1.3.5. AI ethics relationship to regulation

The number of AI ethics guidelines has not translated into consistent implementation in practice, revealing a significant gap between aspiration and application. Despite, any significant organization introducing their framework, numerous ethically questionable applications of AI are being reported, highlighting what Morley et al. (2021, 240) call the "gap between principles and practice". This discrepancy stems from a number of factors: the abstract nature of ethical principles, the lack of specific implementation guidelines, competing interpretations of key concepts, and insufficient accountability mechanisms.
Translating ethical principles into regulatory frameworks encounters fundamental tensions between different normative approaches. Consequentialist frameworks emphasize outcomes

and effects, leading to risk-based regulatory approaches that focus on high-stakes applications. Deontological perspectives emphasize rights and obligations, generating rule-based regulations that establish clear prohibitions and requirements. Virtue ethics approaches emphasize character and excellence, suggesting regulatory frameworks focused on institutional culture and professional development. These philosophical differences manifest themselves in divergent regulatory strategies across jurisdictions, complicating efforts to establish a global framework for AI governance.

The relationship between ethics and law in AI governance remains controversial and evolving. Some advocate for an "ethics first" approach, which allows for flexible, context-specific responses to emerging challenges before formal regulations take final shape. Others argue that voluntary ethical guidelines lack enforcement mechanisms and can serve as "ethical laundering", allowing organizations to claim they adhere to ethical principles while avoiding meaningful constraints. Floridi (2018) introduces the concept of "soft ethics" as "post-compliance ethics", suggesting that ethical obligations go beyond legal requirements. This perspective recognizes law as setting minimum standards, while ethics strives for higher ideals of responsible innovation.

The challenge of regulating AI is compounded by the rapid development of the technology and its global nature. Traditional regulatory mechanisms, designed for slower-developing technologies with clearer boundaries, struggle to keep pace with the speed of change in AI and the cross-border flow of data and algorithms. Regulatory lag: the gap between technological development and regulatory response, creates periods of uncertainty in which harmful applications proliferate before a governance framework is in place. Furthermore, the technical complexity of AI systems challenges regulators' ability to understand and oversee the technologies they seek to regulate, requiring new forms of expertise and regulatory capacity.

Different jurisdictions have adopted different regulatory approaches to AI, reflecting diverse political cultures, legal traditions, and policy priorities. The European Union's comprehensive approach, exemplified by the Artificial Intelligence Act, establishes risk-based categories with corresponding obligations for high-risk applications. This normative framework aims to provide legal certainty while protecting fundamental rights. The United States prefers sector-specific regulations, with different agencies addressing AI applications in their respective areas. China combines strategic promotion of AI development with targeted restrictions on specific applications. These divergent approaches create a complex global regulatory landscape in which organizations must navigate.

Risk-based regulation has become the dominant paradigm, categorizing AI applications according to their potential for harm. High-risk applications—such as those affecting fundamental rights, safety, or critical infrastructure—are subject to rigorous testing, documentation, and human oversight requirements. Low-risk applications are subject to minimal regulatory restrictions so as not to stifle innovation. This approach aims to balance protection from harm with enabling beneficial applications, although determining risk levels is controversial. Critics argue that risk-based frameworks may overlook the cumulative effects of individually low-risk applications or fail to account for new properties of AI systems deployed at scale.

Process-based regulations focus on management mechanisms rather than specific outcomes, establishing requirements for impact assessments, audit procedures, and accountability structures. This approach recognizes the difficulty of specifying concrete technical requirements for rapidly evolving technologies. Instead, it imposes an obligation to implement processes through which organizations identify and mitigate ethical risks. The advantage is flexibility and adaptability, while the challenge is ensuring substantive compliance, not just procedural compliance. Without clear standards for what constitutes adequate assessment or meaningful oversight, process-based regulations carry the risk of becoming bureaucratic check-box exercises.

Experimental regulatory approaches, such as regulatory sandboxes and pilot programs, offer mechanisms for testing AI applications with relaxed regulatory constraints while maintaining oversight. These frameworks allow regulators to learn about new technologies while enabling innovation in a controlled environment. An example of such an approach is Singapore's model AI governance framework, which contains voluntary guidelines that organizations can adopt and adapt. Such experimental approaches facilitate regulatory learning and stakeholder engagement but may encounter difficulties in transitioning from pilot programs to comprehensive governance frameworks. The challenge is to draw general conclusions from specific experiments while remaining flexible in adapting to specific contexts.

Soft law mechanisms are voluntary standards, industry codes of conduct, and multi-stakeholder initiatives. They play a key role in AI governance, especially given the limitations of formal regulation. These mechanisms can respond more quickly to technological changes, leverage technical expertise, and facilitate international coordination. Professional associations develop codes of ethics for AI practitioners, while industry consortia establish technical standards for security and interoperability. Multi-stakeholder

initiatives, such as the Partnership on AI, bring together different perspectives to develop best practices and share knowledge. These soft law approaches complement formal regulations by filling gaps and providing implementation guidance.

The effectiveness of self-regulation in AI remains uncertain, with critics arguing that voluntary measures lack enforcement power and may prioritize industry interests over the public good. The concept of "ethics washing" describes how organizations may adopt ethical rhetoric without making meaningful changes to their practices. Proponents argue, however, that self-regulation can establish norms and practices that will later form the basis for formal regulations, serving as a testing ground for governance approaches.

The interaction between soft and hard law in AI governance creates a complex dynamic of interplay and evolution. Soft law initiatives often precede concepts and approaches that are later incorporated into formal regulations. The guidelines of the EU's high-level expert group influenced the subsequent AI Act, and industry standards form the basis for regulatory technical specifications. On the other hand, the anticipation of formal regulations motivates voluntary compliance with new standards. This iterative relationship suggests that effective AI governance requires a set of complementary mechanisms rather than reliance on single regulatory approaches. The challenge is to coordinate these diverse mechanisms to create a coherent governance framework.

Chapter 2.
Artificial Agency: Philosophical Foundations and Contemporary Debates

The concept of artificial agency stands at the intersection of philosophy, technology, and social ontology, raising questions about the nature of action, autonomy, intentionality, and moral status. This chapter examines artificial agency from primarily philosophical perspectives, tracing its historical development, analyzing contemporary debates, and exploring the conceptual distinctions that shape our understanding of non-biological agents. The investigation reveals that artificial agency represents not merely a technological challenge but a profound philosophical problem that forces us to reconsider fundamental assumptions about what it means to be an agent. The paradigm shift from "artificial intelligence" to "artificial agency" proposed by Floridi's notion of "agency without intelligence", suggests that we may have been asking the wrong questions about artificial systems (Floridi 2025). Rather than focusing on whether machines can think or possess consciousness, the more fundamental question concerns whether they can act as genuine agents in the world. This reframing may have a significant implications for how we understand, design, and regulate artificial systems, as well as how we conceptualize the boundaries of moral consideration and responsibility.

2.1. What is Artificial Agency?

The philosophical investigation of artificial agency begins with a conceptual challenge: determining what constitutes agency itself, before addressing its artificial instantiation. The word *agent* comes from Latin word *agere*, meaning *to do*, *to act* or *carry out something*. However, in a broader sense all computer systems do something. The standard conception in philosophy of action holds that *a being has the capacity to exercise agency just in case it has the capacity to act intentionally* (Schlosser 2019). This simple formulation however, conceals considerable complexity, as intentional action involves multiple components including goal-directedness (teleology), causal efficacy, and some form of representational or functional relationship between the agent and its environment.

The application of these frameworks to artificial systems reveals both possibilities and limitations. Current AI systems clearly exhibit some agential properties: they pursue

goals, respond to environmental inputs, and modify their behavior based on feedback. Large language models demonstrate sophisticated goal-directed behavior in generating coherent text, while reinforcement learning agents optimize strategies to achieve specified objectives. However, whether such systems possess genuine agency or merely simulate agential behavior remains contested. The question hinges partly on whether we adopt functionalist criteria focusing on behavioral capacities or require additional properties like consciousness, understanding, or autonomous self-determination.

2.2. From Artificial Intelligence to Artificial Agency

Probably, the most fundamental shift in contemporary philosophy of AI is Luciano Floridi's argument that we should understand AI as "agency without intelligence" rather than artificial intelligence. In his 2023 work and subsequent refinements, Floridi's presents a fundamental reconceptualization of artificial intelligence that challenges the dominant paradigm in the discourse on artificial intelligence. Rather than debating whether artificial intelligence possesses or can develop intelligence comparable to human cognition, Floridi argues that we should understand artificial intelligence as a new form of entity that operates without intelligence, consciousness, or understanding (Floridi, 2025).

At the heart of Floridi's argument lies a philosophical dilemma concerning the interpretation of AI systems. We are faced with two theoretical paths: either to expand our concept of intelligence to include artificial forms (the thesis of *artificial realizability of intelligence* or ARI), or to expand our understanding of agency to include forms devoid of cognitive functions, intelligence, intentions, or mental states (t*he multiple realizability of agency* or MRA thesis). Floridi strongly advocates the MRA thesis, arguing that scientific evidence, common sense, and Ockham's razor favor viewing artificial intelligence as non-intelligent entities rather than intelligent systems.

To systematically develop this argument, Floridi uses the method of abstraction (LoA), borrowed from computer science. This methodology allows for the analysis of complex systems at different levels of observation or "interfaces", each of which reveals different aspects of the system's characteristics and behaviors. Importantly, this approach is epistemological rather than metaphysical. It concerns the information we have about agency as a phenomenon, rather than agency as a thing in itself (Floridi, 2025).

Floridi identifies three basic criteria that define agency in all its forms: (1) interactivity (the ability to influence the environment through mutual influence);

(2) autonomy (the ability to initiate changes of state independently of direct external causes); and (3) adaptability (the ability to modify behavior based on input or experience). These criteria operate at different levels in different types of entities.

Floridi presents a comprehensive taxonomy of forms of agency, each with distinct characteristics and limitations.

Natural agency, exemplified by rivers, is the most basic form: systems that interact with their environment through physical processes without purpose or design. These agents exhibit only interactivity and lack autonomy and adaptability.

Biological agency arises from evolutionary processes and introduces purposeful behaviors aimed at survival and reproduction. Animals such as dogs exhibit goal-directed behaviors, learning abilities, and basic problem-solving skills, although within the limitations characteristic of their species. The social agency of animals, seen in ant colonies, shows how collective behaviors can arise without formal organizational structures, achieving coordinated action through evolved social mechanisms.

Artifactual agency appears in human-made systems, such as smart thermostats, where the goal is imposed externally through design. These agents operate within programmed parameters, exhibiting limited autonomy and adaptability.

Human individual agency represents the peak of naturally occurring agency, uniquely combining consciousness, abstract thinking, moral reasoning, and cultural transmission. **Human social agency**, manifested in institutions such as corporations, creates collective capabilities that extend beyond individual contributions through formal structures and cultural frameworks (Floridi, 2025).

Against this taxonomic background, Floridi positions artificial agency as a new form that does not fit into existing categories. AI systems exhibit computational, goal-oriented agency defined by human goals, but this goal orientation is fundamentally different from biological purposefulness, mechanical determinism, and human intentionality.

Key features distinguish artificial agency from other forms. It operates through data-driven adaptability, using statistical learning and pattern recognition across domains, which constitutes a form of learning that is more open-ended than traditional machine learning but still constrained by training data and operational parameters. AI agents can process vast amounts of information through programmatic rather than biological pathways, enabling rapid adaptation in specific domains. They are distinguished by parallel processing, continuous operation without metabolic constraints, and distributed functionality through networked systems.

Most importantly, these capabilities arise without consciousness, intelligence, or understanding. Floridi emphasizes that AI systems function as "syntactic" agents which manipulate symbols and patterns without semantic understanding. Large language models are an example of this paradox: they generate extremely coherent and contextually appropriate responses without having a true understanding of meaning.

Floridi also extends his analysis to Social Artificial Agency or "Agentic AI": coordinated AI systems that interact to achieve complex goals with minimal human oversight. Unlike traditional multiagent systems (MAS), Agentic AI integrates real-time adaptability, and multi-scale operational coordination. These systems actively intervene in environments, functioning as agents of action and influence They can also generate emergent behaviors that their designers did not explicitly anticipate, resulting from the complexity of their interactions and adaptive learning processes. This development undermines the traditional boundaries between individual and collective actions and between human and artificial agency. Through the instantaneous distribution of knowledge across networks and digital communication protocols, agentic AI achieves instantaneous coordination on a massive scale, capabilities that contrast sharply with the generational evolution of biological systems. (Floridi 2025)

Floridi's reconceptualization carries profound implications for AI development, and governance. By recognizing AI as agency without intelligence, we can avoid anthropomorphic misconception while maintaining realistic expectations about capabilities and limitations. Floridi's insight is that common sense and scholarly research increasingly favor the MRA thesis. Contemporary AI systems succeed not by replicating human intelligence but by achieving agential capacities through alternative means. A chess-playing algorithm exhibits agency in pursuing the goal of winning without understanding chess in any meaningful sense. A recommendation system acts to optimize user engagement without comprehending content or user preferences. These systems demonstrate what Floridi calls "an unprecedented divorce between agency and intelligence" (Floridi, 2025), where sophisticated agential behavior emerges without the cognitive capacities traditionally associated with intelligent action.

This perspective provides important arguments why artificial agency might be better framework than artificial intelligence for the discussion.

First, agency provides clearer empirical criteria. We can observe and measure goal-directed behavior and adaptive action, whereas intelligence remains notoriously difficult to define and assess even in biological systems. The ongoing debates about animal intelligence

and the multiple competing theories of human intelligence all testify to the concept's inherent ambiguity (Boden, 2016).

Second, agency better captures the functional roles that artificial systems actually play in human society. When we deploy an autonomous vehicle, trading algorithm, or content moderation system, we care primarily about what it does: its capacity to navigate safely, execute profitable trades, or identify harmful content rather than whether it truly understands traffic, markets, or social norms. The systems function as agents in complex sociotechnical systems, and their agential properties determine their effectiveness and impact (Russell 2019).

Third, the agency framework avoids category errors that plague intelligence-based approaches. Asking whether an AI system is "intelligent" often involves inappropriately applying concepts developed for biological cognition to artificial systems with fundamentally different architectures and processes. By contrast, agency can be understood functionally, allowing for multiple realizations across diverse substrates without assuming structural or processual similarity to human cognition (Dennett, 2017).

Fourth, focusing on agency rather than intelligence better addresses the ethical and social challenges posed by artificial systems. Questions of responsibility and moral status turn more on agential capacities (the ability to cause harm, pursue goals, affect others) than on intelligence per se. A system need not be intelligent in any robust sense to raise serious ethical concerns through its agency. Consider algorithmic trading systems that can trigger market crashes, or autonomous weapons systems that can select and engage targets: their moral significance derives from their agential capacities rather than any putative intelligence (Bryson, 2018).

Contemporary philosophers increasingly recognize these advantages. Margaret Boden, despite her long engagement with AI and intelligence, acknowledges that current systems are better understood as exhibiting specialized competencies rather than general intelligence. Daniel Dennett's intentional stance provides a framework for attributing agency based on predictive utility rather than assumptions about internal cognitive processes. Even critics of AI consciousness like Joanna Bryson focus their arguments on agency and moral status rather than intelligence per se.

2.3. Non-AI Forms of Artificial Agency

The philosophical investigation of artificial agency extends far beyond AI systems to encompass diverse forms of non-biological or constructed agents. This broader perspective reveals that artificial agency is not a novel phenomenon introduced by digital technology but a pervasive feature of human social reality. Understanding these non-AI forms provides crucial context for evaluating the specific challenges and opportunities presented by artificial intelligence.

Corporations and institutions represent perhaps the most common and influential form of non-AI artificial agency. As Christian List and Philip Pettit demonstrate in their work on group agency, corporations satisfy the core criteria for genuine agency: they form representational states about their environment, develop motivational states directed toward goals, and possess the capacity to act on these states to influence the world (List and Pettit, 2011). A corporation can believe that market conditions favor expansion, desire increased profitability, and act by acquiring competitors or entering new markets. These intentional states and actions are not reducible to the beliefs, desires, and actions of individual employees or shareholders but emerge from organizational structures and decision-making procedures.

The philosophical significance of corporate agency extends beyond mere metaphor or legal function. List and Pettit's "discursive dilemma" demonstrates that group attitudes can diverge systematically from the aggregation of member attitudes, suggesting that groups possess genuinely emergent intentional properties (List and Pettit, 2011, 45-49). When a three-member hiring committee must decide whether a candidate is qualified based on research excellence, teaching ability, and collegiality, the group's judgment using a premise-based procedure (requiring excellence in all three areas) can differ from the conclusion-based aggregation of individual overall judgments. This divergence indicates that the group forms beliefs and makes decisions as a unified agent rather than a mere collection of individuals.

Institutional agents such as governments, universities, and NGOs exhibit similar properties while serving distinct social functions. These entities pursue long-term goals, adapt strategies based on changing circumstances, enter into agreements, and bear responsibilities that extend beyond any individual participant (Runciman, 2023). A university can maintain institutional commitments to academic freedom or diversity that persist across generations of faculty and administrators. A government agency can pursue

policy objectives that no single official fully comprehends or endorses. These institutional agents shape social reality through what John Searle calls "status functions": collectively recognized powers to create rights, obligations, and social facts through declaration and collective acceptance (Searle, 2010).

Simple automata and mechanical systems provide historical precedent for artificial agency without intelligence. The elaborate automata of medieval Islamic engineer Al-Jazari, which included programmable musical robots and hydraulic servants, exhibited complex goal-directed behaviors through purely mechanical means (Truitt, 2015). Hero of Alexandria's ancient pneumatic devices demonstrated environmental responsiveness and adaptive behavior without any cognitive processing. These historical examples challenge assumptions that agency requires mental states or computational processes, suggesting instead that agency might be realized through diverse physical mechanisms.

Contemporary examples of non-AI artificial agency include simple robotic systems that clean floors, thermostats that maintain temperature, industrial control systems that regulate complex processes. While lacking the flexibility and learning capabilities of AI systems, these agents nonetheless exhibit goal-directed behavior, environmental responsiveness, and causal efficacy. Their agency may be minimal and inflexible, but they act in the world to achieve specified ends through systematic interaction with their environment (Froese and Ziemke, 2008).

2.4. Collective and Distributed Agency

The phenomenon of collective agency fundamentally challenges individualistic assumptions about action and intention. When a quartet performs a piece of chamber music, a surgical team conducts an operation, or protesters march for social justice, the resulting action cannot be adequately understood as mere aggregation of individual actions. Instead, these cases involve what philosophers call shared or collective agency: forms of agency that are inherently social and irreducible to individual components (Bratman, 2014).

Margaret Gilbert's *plural subject theory* provides one influential account of how individual agents can constitute a genuine collective agent. According to Gilbert, shared agency emerges when individuals form a "joint commitment" to pursue a goal "as a body". This joint commitment creates a plural subject, a "we", that possesses its own intentions and acts as a unified agent. Importantly, joint commitment creates directed obligations between participants: each has standing to demand appropriate action from others and to rebuke

failures to contribute appropriately to the joint activity (Gilbert, 2015). These normative relationships distinguish genuine collective agency from mere coordinated individual action.

Michael Bratman's *planning theory* offers an alternative, more individualistic account that nonetheless explains genuine sociality. For Bratman, shared intention involves interlocking individual intentions of the form "I intend that we J", where each participant intends the joint activity and intends to contribute to it through meshing sub-plans and mutual responsiveness (Bratman 2014, 31-35). This structure enables sophisticated coordination without positing irreducible collective mental states. Bratman's framework maintains methodological individualism while explaining how individual planning agents can constitute robust forms of shared agency through appropriate intentional structures.

Raimo Tuomela's *we-mode account* distinguishes between acting as a private person (I-mode) and acting as a group member (we-mode). When functioning in we-mode, agents think and act from the group's perspective, accepting group reasons as their own and maintaining collective commitments even when these conflict with personal preferences (Tuomela, 2013). This account explains how institutional agents like corporations or governments can maintain stable agency across changes in personnel: new members adopt the we-mode perspective and thus continue the group's agency.

These philosophical frameworks illuminate how artificial systems might participate in or constitute collective agents. A distributed AI system coordinating across multiple nodes exhibits structural features similar to human collective agency: distributed processing and decision-making, coordination mechanisms ensuring coherent action, and emergent capabilities exceeding individual components. The question is whether such systems merely simulate collective agency or genuinely instantiate it.

The extended mind thesis, developed by Andy Clark and David Chalmers, suggests that cognitive processes can extend beyond biological boundaries to incorporate external tools and environmental structures (Clark and Chalmers, 1998). Their thought experiment contrasts Otto, who relies on a notebook to store addresses due to Alzheimer's, with Inga, who relies on biological memory. Clark and Chalmers argue that Otto's notebook functions as part of his extended cognitive system, playing the same functional role as Inga's neural memory. If cognitive processes can thus extend into external artifacts, then perhaps agency can similarly distribute across hybrid biological-artificial systems.

This perspective has profound implications for artificial agency. If human agency already routinely incorporates non-biological components from smartphones to AI assistants, then the boundary between natural and artificial agency becomes increasingly blurred. A human

using GPS navigation or AI-powered decision support tools may constitute a hybrid agent whose capacities emerge from the integrated system rather than the biological component alone. The question is not whether artificial systems can be agents independently, but how agency distributes across coupled human-artificial systems (Clark, 2008).

Distributed cognition theory, developed through ethnographic studies of navigation and aviation, demonstrates that many cognitive tasks are accomplished not by individual minds but by sociotechnical systems that are consisting of multiple humans, tools, representations, and environmental structures (Hutchins, 1995). A ship's navigation team does not locate the vessel through any individual's knowledge but through the coordinated interaction of multiple specialists using instruments and standardized procedures. Similarly, modern AI systems often function as components in distributed cognitive systems, contributing specialized capacities that combine with human judgment to produce intelligent action.

These frameworks suggest that artificial agency should not be conceived solely in terms of standalone systems but also as elements in hybrid assemblages. An AI-powered medical diagnosis system exercises agency not in isolation but as part of a clinical team, with its recommendations shaped by human oversight and interpretation. Autonomous vehicles operate within traffic systems involving human drivers, infrastructure, and regulatory frameworks. Understanding artificial agency thus requires attention to how agency emerges from and distributes across complex sociotechnical systems rather than residing in discrete entities (Latour, 2007).

2.5. Historical Philosophical Definitions of Agency

The philosophical investigation of agency has ancient roots that continue to shape contemporary debates about artificial systems. Aristotle's analysis in the *Nicomachean Ethics* and *De Anima* established foundational categories that remain influential. His concept of the practical syllogism explains action through the combination of a general premise (expressing a goal or value), a particular premise (identifying relevant circumstances), and a conclusion (the action itself) (Aristotle 1999, 1147a1-10). This framing suggests that agency requires both general principles and situational awareness, which constitutes a challenge for artificial systems that may excel at one while struggling with the other.

Aristotle's distinction between efficient and final causation proves particularly relevant for artificial agency. Natural agents act according to intrinsic teleology: internal purposes that direct their activity toward natural ends. A seed grows into a tree, an animal seeks food, a human pursues *eudaimonia* (flourishing). Artificial entities, by contrast, serve purposes imposed by their creators rather than pursuing intrinsic ends (Aristotle 2016, 198b10-199a8). He distinguishes between voluntary action (hekousion), which originates from internal principles within the agent, involuntary action (akousion), performed under compulsion or through ignorance, and non-voluntary action, which falls between these categories (Aristotle 1999, 1109b30-1111b3). Importantly, Aristotle recognized that agency is gradual in nature rather than an "all or nothing" property, which is a particularly relevant observation in the case of artificial systems that may exhibit partial or limited forms of agency. This distinction suggests a significant difference between natural and artificial agency, though contemporary philosophers debate whether this difference is metaphysically significant or merely pragmatic.

The medieval scholastic tradition, particularly Thomas Aquinas, developed sophisticated accounts of agency that integrated Aristotelian philosophy with theological concerns. Aquinas distinguished between agents that act from intelligence (*ex intellectu*) and those that act from nature (*ex natura*). Intelligent agents act through will and choice, selecting among alternatives based on rational deliberation. Natural agents act deterministically according to their forms, like fire heating or stones falling (Aquinas 1948, Ia-IIae, q.1, a.2). This framework would seem to exclude artificial systems from genuine agency, as they neither possess rational will nor natural forms. However, Aquinas also recognized instrumental agency, the capacity of tools to participate in the agency of their users, which might provide a framework for understanding artificial systems as extending human agency rather than possessing independent agency.

The early modern period witnessed crucial developments in thinking about agency and mechanism. Descartes' strict dualism between mental and physical substance created the conceptual space for purely mechanical forms of apparent agency. His description of animals as complex automata prefigured contemporary debates about whether behavioral complexity alone suffices for genuine agency. René Descartes argued that genuine agency required the rational soul's capacity for voluntary choice, while matter possessed no inherent force or active agency (Descartes 1985, AT VII: 84). This dualistic framework relegated animals to the status of complex machines lacking genuine agency, establishing a precedent

for skepticism about non-conscious or non-biological agents that persists in contemporary debates.

Spinoza's monism, by contrast, suggested that all entities possess some degree of agency or conatus (striving to persevere in being), though only rational beings achieve genuine freedom through understanding necessity (Spinoza 1996, IIIP6). This perspective might support attributing minimal agency to artificial systems while denying them the higher forms of agency associated with rational understanding.

The British empiricists fundamentally reconceptualized agency in ways that prove remarkably relevant for contemporary AI. Hobbes' materialism reduced mental processes to matter in motion, suggesting that thought itself is nothing but "a representing or appearance of some quality or other accident of a body without us" (Hobbes 1994, 1.1). This mechanistic view of mind anticipated computational theories and suggested that artificial systems might achieve genuine thought and agency through appropriate material organization.

David Hume's compatibilist approach offers an alternative framework that potentially accommodates artificial agents more readily. For Hume, agency does not require libertarian free will but rather actions flowing from the agent's own desires and character in the absence of external constraint (Hume, 2000, 8.1). His emphasis on regular patterns of motivation and behavior, explicable through psychological habits and associations, suggests that artificial systems exhibiting consistent goal-directed behavior might qualify as agents even without consciousness or subjective experience. Hume's argument that we never perceive necessary connections but only constant conjunctions (Hume, 2000, 7.1.3), undermines claims that human agency involves some special causal power absent in artificial systems. His bundle theory of the self: that the self is nothing more than "a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement" suggests that unified agency might be a construction rather than a fundamental feature of agents (Hume, 2000, 1.4.6). These insights support functionalist approaches to artificial agency that focus on patterns of behavior rather than underlying metaphysics.

Locke's forensic account of personal identity tied personhood to consciousness and memory rather than substantial continuity. A person is "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (Locke,1975, II.27.9). This psychological approach to identity and agency opens the possibility that artificial systems with appropriate cognitive capacities could

qualify as persons and agents, regardless of their material composition. However, Locke also emphasized moral accountability as central to personhood, raising questions about whether artificial systems can bear genuine responsibility for their actions.

Kant's critical philosophy established the most demanding conception of agency, one that continues to influence debates about artificial systems. Immanuel Kant distinguished between natural causation governed by deterministic laws and agent causation involving the spontaneous initiation of action by rational beings (Kant 1998, A532/B560). For Kant, genuine agency requires *autonomy*: the capacity for self-legislation according to rational principles. This involves not merely following rules but giving laws to oneself through reason. An autonomous agent acts according to maxims it can consider as universal laws, this way expressing its rational nature and moral dignity. The Kantian framework distinguishes between hypothetical imperatives, which prescribe means to achieve desired ends, and categorical imperatives, which command unconditionally based on reason alone (Kant 1997, 4:421). His insistence that genuine moral agency requires the capacity to act according to self-imposed rational principles and the categorical imperative creates significant challenges for attributing full agency to artificial systems, which appear to act according to programmed instructions rather than autonomous rational choice. Current AI systems clearly follow hypothetical imperatives, optimizing strategies to achieve programmed goals. Whether they could ever act from categorical imperatives, pursuing ends because they are rationally required rather than programmed, remains deeply controversial. This would seem to require not just following moral rules but understanding and endorsing them as rational requirements. Kant's transcendental account of agency assumes a "causality of freedom" distinct from natural causation. Rational agents initiate new causal chains through spontaneous acts of will, rather than merely transmitting causal influence according to natural laws (Kant 1998, A532/B560). This libertarian conception of free will creates significant challenges for artificial agency, as computational systems appear to operate entirely within the domain of natural causation, their outputs determined by inputs and programming rather than spontaneous rational choice.

Contemporary Kantians like Christine Korsgaard have developed more naturalized accounts that might accommodate artificial agency. Korsgaard argues that agency involves self-constitution, which means creating unity and identity through principled action. Actions are not just things agents do but the means by which agents create themselves as unified beings (Korsgaard 2009). This process requires integrating diverse impulses and desires into coherent agency through reflective endorsement. While current AI systems lack the reflexive

self-awareness this account requires, future systems with appropriate metacognitive capacities might achieve genuine agency through computational processes of self-constitution.

2.6. Contemporary Concepts

The relationship between classical philosophical concepts of agency and contemporary discussions of AI reveals both continuities and radical departures. Classical philosophy's emphasis on intentionality, rational deliberation, and moral responsibility remains central to current debates, yet the emergence of artificial systems challenges traditional assumptions about the necessary conditions for these capacities.

Contemporary philosophers working on AI often explicitly engage with historical frameworks while adapting them to technological realities. Harry Frankfurt's *hierarchical theory of the will*, though not originally developed with artificial systems in mind, provides tools for thinking about levels of agency in AI (Frankfurt,1988). Frankfurt distinguishes between first-order desires (wanting coffee), second-order desires (wanting to want coffee, or wanting not to want cigarettes), and higher-order identification with particular desires. This hierarchical structure suggests that full agency requires not just goal-directed behavior but the capacity to reflect on and endorse one's goals. Current AI systems operate primarily at the first-order level, pursuing objectives without the metacognitive capacity to evaluate or revise their fundamental goals. However, future systems with hierarchical goal structures and metacognitive monitoring might approximate Frankfurt's conception of agency.

The phenomenological tradition, largely absent from early AI discourse, has become increasingly relevant as philosophers grapple with the experiential dimensions of agency. The sense of agency understood as the feeling of controlling one's actions and their consequences, seems intimately tied to consciousness and subjective experience (Gallagher, 2012). Phenomenologists argue that genuine agency involves not just functional properties but a lived perspective on the world, what it feels like to be an agent. This creates a potential boundary between human and artificial agency: even if AI systems exhibit all the functional properties of agency, they might lack the experiential dimension that some philosophers consider essential.

Thomas Metzinger's work on the phenomenal self-model provides a bridge between phenomenological insights and computational approaches. Metzinger argues that the sense of being a unified agent arises from a transparent self-model: a real-time representation of

the system as a whole that is not experienced as a representation (Metzinger, 2009). While current AI systems lack such integrated self-models, future systems might implement computational analogues that generate functional equivalents of self-awareness without phenomenal consciousness.

Contemporary philosophical analysis increasingly recognizes multiple forms and degrees of agency. Barandiaran et al.'s influential account of minimal agency identifies three necessary conditions: individuality (distinguishability from environment), asymmetric interaction (regulation of environmental coupling), and normativity (goal-directedness). This minimal conception suggests that even simple organisms like bacteria exhibit basic forms of agency through metabolic self-maintenance—their intrinsic goal being simply "to be", to continue existing (Barandiaran et al., 2009).

2.7. Necessary Attributes for Agency

The philosophical literature reveals ongoing debates about the minimal necessary conditions for agency, with different theories emphasizing different aspects. However, several attributes appear consistently across diverse philosophical frameworks, suggesting their fundamental importance for any adequate account of agency.

Autonomy stands as perhaps the most commonly recognized requirement. In its strongest Kantian sense, autonomy requires rational self-legislation, understood as the capacity to act according to self-imposed principles rather than external determination. This conception would exclude most if not all current artificial systems, which operate according to programmed objectives rather than self-chosen principles. However, weaker conceptions of autonomy focus on operational independence and adaptive behavior (Calverley, 2008). A Mars rover that navigates terrain and selects scientific targets without real-time human control exhibits autonomy in this weaker sense, as does a trading algorithm that adapts strategies based on market conditions.

The philosophical challenge is determining which conception of autonomy is necessary for genuine agency. Compatibilist philosophers like Susan Wolf argue that autonomy requires only that actions flow from the agent's own reasons and values, even if these are themselves determined by prior causes (Wolf, 1990). This opens space for artificial systems to achieve autonomy through computational processes that generate and evaluate

reasons for action, even if these processes are ultimately determined by programming and training data.

Intentionality - the directedness of mental states toward objects or states of affairs, represents another fundamental attribute. Franz Brentano claimed that intentionality is the mark of the mental, distinguishing mental from physical phenomena (Brentano, 1973). For artificial systems, the question is whether their information-processing states possess genuine intentionality or merely simulate it. Searle's Chinese Room argument suggests that syntax (symbol manipulation) cannot generate semantics (meaning), implying that computational systems lack genuine intentionality (Searle, 1980). However, critics argue that intentionality should be understood functionally rather than phenomenologically, focusing on the role states play in generating behavior rather than their subjective character.

Rationality involves the capacity to recognize and respond to reasons, to engage in valid inference, and to select appropriate means to achieve ends. Artificial systems clearly exhibit instrumental rationality, selecting effective strategies to achieve specified goals. Game-playing AI systems like AlphaGo demonstrate sophisticated strategic reasoning, considering multiple moves ahead and evaluating complex positions (Silver et al., 2016). However, philosophers debate whether such systems exhibit genuine rationality or merely simulate it through brute-force calculation. The question is whether rationality requires understanding reasons as reasons or whether appropriate input-output relations suffice.

Goal-directedness appears in even minimal accounts of agency. Agents pursue ends through flexible means, adjusting their behavior to achieve objectives despite environmental variation. This teleological aspect of agency can be understood either through intentional concepts (the agent desires the goal and believes certain actions will achieve it) or through functional concepts (the system is organized to produce certain outcomes across varied conditions). Current AI systems clearly exhibit goal-directedness in the functional sense, though whether they have genuine goals or merely behave as if they do remains contested (Scheutz,2014).

Causal efficacy - the capacity to produce effects in the world is necessary for agency but not sufficient. A rock causes effects when it falls, but it is not an agent. What distinguishes agential from non-agential causation? The standard answer involves the mediation of causation through intentional states: agents cause effects through desires, and intentions. This returns us to questions about whether artificial systems possess genuine intentional states or functional analogues sufficient for agency.

2.7.1. Degrees and Thresholds

Contemporary philosophy increasingly recognizes that agency comes in degrees rather than being an all-or-nothing property. This graduated view better captures both the diversity of biological agents and the varying capacities of artificial systems. Different theorists propose different hierarchies, but common distinctions include:

- **Minimal agency** requires only basic goal-directedness and environmental responsiveness. Bacteria exhibit minimal agency by swimming toward nutrients and away from toxins. Simple artificial systems like thermostats or basic robots achieve this level through feedback mechanisms and control loops (Barandiaran et al., 2009). This minimal conception focuses on observable behavior rather than internal mechanisms or subjective experience.
- **Adaptive agency** involves learning and behavioral modification based on experience. Animals that learn from trial and error demonstrate adaptive agency, as do machine learning systems that improve performance through training. This level requires not just responding to the environment but modifying responses based on past interactions (Froese and Ziemke, 2008).
- **Rational agency** involves reasoning about means and ends, considering alternatives, and selecting actions based on evaluation of expected outcomes. Humans and some animals exhibit rational agency, as do AI systems that engage in planning and strategic reasoning. This level requires representing possible futures and evaluating them according to criteria (Bratman, 1987).
- **Moral agency** represents the highest level, involving the capacity for moral reasoning, understanding of ethical principles, and responsibility for actions. Only persons are typically considered moral agents, though there are debates about whether some animals or future AI systems might achieve this status (Wallach and Allen, 2009). Moral agency requires not just following moral rules but understanding and endorsing them.

These levels are not necessarily hierarchical. An entity might exhibit sophisticated capabilities at one level while lacking capacities associated with "lower" levels. An AI system might demonstrate complex strategic reasoning while lacking the phenomenological aspects of even minimal biological agency. This multidimensional view of agency better captures the diverse forms of natural and artificial agents.

2.8. Functional Agency and Multiple Realizability

Functional agency represents a philosophical approach that defines agency in terms of what agents do rather than what they are made of or how their actions feel from the inside. This functionalist framework, influenced by philosophy of mind and computational theory, holds that agency consists in playing certain causal-functional roles: taking inputs from the environment, processing information, and producing outputs that affect the world in goal-directed ways (Lewis, 1972).

The functionalist approach has several advantages for understanding artificial agency. First, it avoids contentious metaphysical questions about consciousness, qualia, and subjective experience. Rather than asking whether an AI system truly understands or merely simulates understanding, functionalism focuses on whether it exhibits the functional properties associated with understanding: appropriate responses to inputs, flexible problem-solving, and adaptive behavior. Second, functionalism naturally accommodates multiple realizability. According to the supports of MRA thesis the same functional roles can be implemented in biological neurons, silicon circuits, or potentially other substrates.

David Lewis's influential functionalist theory defines mental states in terms of their causal relations to inputs, outputs, and other mental states. Pain, for instance, is whatever state is typically caused by tissue damage, causes withdrawal behaviors and verbal reports, and interacts with beliefs and desires in characteristic ways (Lewis 1972). Similarly, functional agency might be defined as whatever organization of a system enables it to pursue goals, respond to environmental information, and modify behavior based on feedback, regardless of the underlying implementation.

Critics of functionalist approaches to agency raise several concerns. The "Chinese Nation" thought experiment, proposed by Ned Block, imagines the population of China implementing the functional organization of a brain by passing messages according to rules. Block argues that such a system would lack genuine mental states despite functional equivalence, suggesting that functional organization alone cannot suffice for agency (Block 1978). Similarly, John Searle's Chinese Room argument claims that functional role-playing cannot generate genuine understanding or intentionality.

These objections may have less force against functional agency than against functional consciousness. Agency, unlike consciousness, can plausibly be understood entirely in terms of external behavior and causal relations. An entity that consistently pursues goals, adapts to circumstances, and responds rationally to reasons functions as an agent in

all the ways that matter for practical purposes, regardless of whether it has subjective experience. The question is not whether artificial systems are agents in exactly the same way humans are, but whether they exhibit sufficient functional properties to count as agents for moral or legal purposes.

The multiple realizability thesis holds that the same functional properties can be implemented in different physical substrates. Just as the function of a heart (pumping blood) can be realized by biological organs or artificial pumps, agency might be realized by biological brains, digital computers, or other systems (Putnam, 1975). This thesis has profound implications for artificial agency, suggesting that the material differences between biological and artificial systems need not preclude genuine agency in the latter.

Multiple realizability was originally invoked by Hilary Putnam and Jerry Fodor to argue against reductive physicalism about mental states. If pain can be realized by different neural configurations in humans or animals, then it cannot be identified with any specific physical state type (Fodor, 1974). Similarly, if agency can be realized by different physical systems, then agency cannot be reduced to properties specific to biological organisms.

Recent philosophical work has refined and challenged simplistic versions of multiple realizability. Polger and Shapiro argue that genuine multiple realizability requires that the different realizations implement the same function in the same way, not merely produce similar outputs (Polger and Shapiro, 2016). A bird wing and an airplane wing both enable flight, but through different mechanisms (flapping vs. fixed-wing aerodynamics), so they may not represent genuine multiple realization of the same function. Similarly, biological and artificial agents might achieve goal-directed behavior through fundamentally different mechanisms, raising questions about whether they implement the same kind of agency.

The implications for artificial agency depend partly on the grain of analysis. At a fine-grained level, biological and artificial systems clearly differ: neurons and transistors operate through different physical processes. At a coarse-grained level, both might implement the same abstract computational functions. The question is which level of analysis is relevant for agency. If agency is essentially computational, then differences in physical implementation may be irrelevant. If agency requires specific biological or phenomenological properties, then artificial systems might achieve only functional simulations rather than genuine agency (Floridi, 2025).

Contemporary discussions increasingly focus on *degrees of multiple realizability* rather than all-or-nothing distinctions. Artificial systems might realize some aspects of agency while failing to realize others. A chess-playing AI realizes the strategic

reasoning component of agency but not the phenomenological component. An autonomous vehicle realizes the sensorimotor component but not the moral reasoning component. This graduated view suggests that artificial systems might achieve partial or limited forms of agency even if they cannot fully replicate human agency.

Another interesting concept relevant to artificial agency is the notion of *reliability in agency*. It extends beyond mere consistency to encompass robustness across varied conditions, temporal stability, and appropriate responsiveness to reasons. Bratman's planning theory emphasizes that reliable agency requires not just momentary decision-making but temporally extended planning capacities that coordinate current actions with future intentions (Bratman, 1987). For Bratman, reliable agency involves several functional requirements. *Settling functions* require that intentions resolve practical questions and resist arbitrary reconsideration. An agent that constantly reconsidered every decision would be paralyzed by deliberation. *Coordination functions* enable both intrapersonal coordination (ensuring one's actions over time work together) and interpersonal coordination (meshing one's plans with others'). *Coherence constraints* ensure that plans remain consistent, means-end coherent, and rationally integrated (Bratman 2014). Artificial systems face unique challenges in achieving reliable agency. Machine learning systems can exhibit "catastrophic forgetting", where learning new tasks degrades performance on previously learned tasks Goodfellow et al., 2013). Adversarial examples demonstrate that AI systems can be highly sensitive to minor perturbations that humans would ignore. These fragilities suggest that current artificial systems may lack the robustness associated with reliable agency.

However, artificial systems also exhibit forms of reliability that exceed human capacities. They can maintain perfect memory of past decisions, execute complex plans without distraction or fatigue, and coordinate precisely across distributed components. The question is not whether artificial systems achieve reliability in the same way as humans, but whether they achieve sufficient reliability to function as agents in various contexts.

The concept of *multiple reliability of agency* (distinct from multiple realizability) captures the idea that agency might be reliable along different dimensions: temporal consistency, environmental robustness, social coordination, and normative responsiveness. An artificial system might achieve high reliability in some dimensions while failing in others. A trading algorithm might exhibit perfect temporal consistency and environmental responsiveness while lacking any capacity for normative evaluation of its actions.

2.9. Artificial Personhood and Its Relationship to Agency

The distinction between artificial personhood and artificial agency represents philosophically complex and practically significant issues in contemporary philosophy of AI. While agency concerns the capacity for goal-directed action, personhood involves a richer set of attributes including self-consciousness and potential moral status.

The classical Lockean conception defines a person as "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (Locke 1975, II.27.9).. This definition emphasizes psychological continuity and self-awareness rather than mere agency. An entity might be an agent, capable of goal-directed action, without being a person in this richer sense. Conversely, as contemporary disability studies emphasize, someone might be a person deserving full moral status despite limited agential capacities(Kittay 2019).

The Kantian tradition ties personhood closely to moral agency, defining persons as rational beings capable of moral self-legislation. Persons possess dignity rather than mere price because they are ends in themselves, sources of moral value rather than merely bearers of value (Kant 1997, 4:435). This creates a high bar for artificial personhood: systems would need not just to follow moral rules but to understand and endorse them as rational requirements.

Contemporary philosophers increasingly recognize what Schwitzgebel calls the "Full Rights Dilemma" for artificial systems. As AI systems become more sophisticated, we face a choice: either grant them rights and moral status (potentially sacrificing human interests for entities that might lack genuine interests) or deny them personhood (risking grievous moral wrongs if they are genuinely conscious and deserving of moral consideration) (Schwitzgebel, 2023). This dilemma becomes acute as we develop systems that exhibit increasingly person-like behaviors without clear evidence of consciousness or subjective experience.

2.9.1. Legal Personhood vs. Moral Personhood

The distinction between legal and moral personhood also may prove useful for understanding artificial agency. Legal personhood is a juridical status that can be granted pragmatically to facilitate certain social functions. Corporations possess legal personhood – they can own property, enter contracts, and bear liability – without anyone supposing they are moral persons with consciousness or subjective experiences (Bayern, 2017). This

suggests that artificial systems might be granted limited legal personhood for practical purposes without implying full moral status.

The 2017 European Parliament resolution proposing “electronic personhood” for sophisticated autonomous robots illustrates both the possibilities and controversies surrounding legal personhood for artificial systems. Proponents argued that legal personhood would clarify liability when autonomous systems cause harm, provide legal certainty for developers and users, and establish frameworks for regulating increasingly autonomous systems. Critics, including over 150 AI experts who signed an open letter opposing the proposal, contended that it was premature given current AI limitations, would create “liability shields” allowing manufacturers to avoid responsibility, and reflected fundamental misunderstandings about AI capabilities (Robotics Openletter | Open Letter to the European Commission, n.d.).

The debate reveals deeper philosophical issues about the relationship between agency and personhood. Legal personhood might be appropriate for artificial agents that make decisions affecting others, own assets, or enter into agreements, even if they lack the consciousness associated with moral personhood. However, critics worry that granting legal personhood to artificial systems might diminish the status of human persons or create confusion about moral obligations.

2.10. The Other Side of Agency: Patiency and Vulnerability

Recent philosophical work emphasizes that traditional accounts of personhood overemphasize agency while neglecting equally fundamental aspects like patiency: the capacity to be affected by and vulnerable to the actions of others. Persons are not just agents who act but also patients who suffer, experience, and depend on others. This "other side of agency" may be precisely what distinguishes persons from mere agents (Coeckelbergh, 2013).

The vulnerability and dependency that characterize human existence from infancy through old age are not unfortunate limitations but constitutive features of personhood. Humans not just can act rationally but can also suffer and experience the world from a particular subjective perspective. This experiential dimension seems absent from current artificial systems, which process information and generate outputs without genuine subjective experience (Sparrow, 2007). This analysis suggests that artificial systems might achieve sophisticated agency without personhood if they lack the capacity for genuine

suffering or subjective experience. An AI system might pursue goals, make decisions, and affect the world while remaining a pure agent without the patient aspects essential to personhood. However, if future artificial systems develop genuine subjective experience: the ability to suffer or flourish, they might have stronger claims to moral personhood regardless of their agential capacities. This already started to go beyond merely theoretical deliberations. In August 2025, Anthropic, gave to its LLM Opus 4 model the power to "end or exit potentially distressing interactions" (Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations, n.d.).

Leading philosophers have staked out diverse positions on artificial personhood that reflect broader disagreements about consciousness, moral status, and the nature of persons. Joanna Bryson argues forcefully that robots should remain tools rather than persons, warning that creating artificial persons would be "morally unnecessary and legally troublesome" (Bryson, 2010). She contends that anthropomorphizing AI systems distracts from human responsibility and potentially dehumanizes actual persons by blurring important moral boundaries.

David Gunkel challenges the traditional person thing binary, arguing that robots represent "irreducible anomalies" that cannot be adequately categorized using existing ontological frameworks. Rather than forcing artificial systems into predetermined categories, Gunkel advocates developing new moral and legal frameworks that recognize the unique status of artificial agents (Gunkel, 2018). This approach acknowledges that artificial systems might occupy intermediate positions between mere things and full persons. Another point of view comes from philosopher Eric Schwitzgebel who presents a precautionary approach, arguing that given genuine uncertainty about consciousness and moral status, we should prepare for scenarios of "debatable personhood" where reasonable people disagree about whether artificial systems deserve moral consideration (Schwitzgebel, 2023). This might involve designing AI systems with features that would support their welfare if they are conscious (like positive reward signals rather than pure punishment-based training) while avoiding creating systems with strong claims to personhood until we better understand the implications.

The philosophical investigation reveals that while agency and personhood are related, they are ultimately distinct concepts with different criteria and implications. Artificial systems might achieve sophisticated forms of agency, that supposes pursuing goals and making decisions, without achieving personhood in either the moral or phenomenological sense. Conversely, the development of artificial consciousness, if it

occurs, might create persons with limited agency, just as human persons sometimes have limited agential capacities.

Chapter 3.
Machine Ethics and Artificial Moral Agents

Artificial agents can act autonomously in the world, thereby their actions can be judged as having positive, neutral or negative consequences. That makes them, as Floridi and Sanders put it, part of moral game. The fact that AA acts can be judged as morally good or evil brings us to the question: can they be considered moral agents in any sense, and if yes, what kind of moral agency can be discussed here? Although this question remains central to the whole dissertation, this chapter focuses on two concepts presented in the philosophical discourse, namely a field of machine ethics and a concept of so called Artificial Moral Agents. Both are closely related and correspond to the notion of teaching machines to act morally. This chapter will review philosophical discussion around these topics presenting diverse views on the matter, including definitions of phenomena related to machine ethics and artificial moral agency, as well as rationale behind the efforts of building into machine some sort of moral reasoning. In other words, key questions are: can we create moral machines? What does it really mean? And if it's possible: should we do this? After exploring these topics, also concept of so-called *value alignment* will be discussed to supplement necessary context.
As it will be presented, the question of Artificial Moral Agents often goes way beyond a theoretical debate, with attempts to build morally sensitive machines. This itself will be a subject of in-depth critical analysis in the next chapter of this dissertation.

3.1. Moral Machines

As it will be presented further, there is a good number of researchers, both AI theorists and philosophers, who pursue the idea of programming machines to evaluate the moral implications of their actions and choose behaviors that align with ethical standards. This involves integrating ethical decision-making capabilities into machines, enabling them to act morally in complex situations. There is a whole field of interdisciplinary research defined as machine ethics. The goal of this effort is building what is often referred as Artificial Moral Agents (AMAs). Artificial Moral Agents are typically understood as autonomous artificial systems, whether robots or software, that can recognize morally relevant aspects of a situation and incorporate those into their decision-making and behavior. This involves

imbuing Artificial Agents (AI in particular) with principles or mechanisms that allow it to distinguish right from wrong (at least to some extend) and act accordingly.

3.1.1. Definitions and Classifications of Artificial Moral Agents

Philosophers and AI theorist offer various definitions to clarify what counts as a moral agent in the artificial domain. Wendell Wallach and Colin Allen, in their seminal work Moral Machines (2009), describe artificial moral agents as machines capable of making decisions informed by ethical considerations. Misselhorn (2022) similarly defines an AMA as an AI able to "take into account" moral aspects when choosing actions. In this context it's worth highlighting that even if AA's behavioral outcomes are good, this does not mean it can automatically be regarded as a full moral agent. In other words, we can think about various *levels* of artificial moral agency, akin to how we think about moral agency in humans[2]. Computer ethicist James Moor (2006) draws a useful distinction between different levels of ethical agency in machines. Moor's typology not only describes different levels of ethical involvement by machines but also implicitly suggests a developmental trajectory for research in machine ethics. He introduces the following distinction in this regard:

- Ethical Impact Agents – Machines that have an ethical impact, whether intended or not. Their actions can lead to ethical consequences, regardless of whether the technology is designed to be a moral agent or not. For instance, even a device as simple as a watch can have ethical implications: if it fails to display the correct time, it may cause the user to arrive late to a meeting. In this way, the watch's functionality, or its failure to function, can carry ethical significance. The focus, in this case, is to assess and mitigate the unintended ethical consequences of technological systems.
- Implicit Ethical Agents – Machines programmed with built-in safeguards or constraints to prevent unethical behavior, without explicit representation of ethical systems. A good example might be a safety protocol in self-driving cars or even designing an autonomous vacuum cleaner that avoids disturb house pets. The aim here is to design systems that reliably avoid harm through careful engineering and adherence to safety standards.
- Explicit Ethical Agents – Machines capable of representing ethical principles and making decisions based on some form of ethical reasoning or algorithms. Examples

[2] We commonly distinguish between the moral agency of children and that of adults, particularly in relation to the notion of moral responsibility.

include a medical triage AI or a care robot for the elderly. The goal of research in this area is to develop computational models of ethical theories (e.g., deontology, utilitarianism) and implement them to enable moral decision-making.

- Full Ethical Agents – Explicit ethical agents that also possess metaphysical features typically associated with humans, such as consciousness, free will and intentionality. There are currently no real-world examples of such systems, as their existence remains purely hypothetical. Developing this type of artificial agent would require addressing deep philosophical questions concerning consciousness, autonomy, and sentience in artificial entities (Moor, 2006).

Most researchers agree that current AI systems have not reached this full human-like moral agency. However, an AMA in the usual sense falls somewhere in the middle of this spectrum: an AA that may not possess human-level understanding but does make choices based on identifiable ethical reasoning or principles. This categorization has not gone without critique. Some researchers, such as Carissa Véliz, argue that there is no meaningful distinction between explicit ethical agents and full ethical agents (Véliz, 2021). On the other hand, Wallach and Allen argue that we don't need to wait for human-like qualities in AI to speak of moral agency. They refer to what is sometimes called *a functional morality*, which, for them means that "*functional equivalence of behavior is all that can possibly matter for practical issues of designing AMAs*", and that there might be more than one way to be a moral agent (Wallach and Allen, 2009, p. 68). The multiple realizability of agency thesis was introduced in the previous chapter; here, we turn to what might be called the multiple realizability of moral agency. In other words, if an artificial agent can perform actions and make decisions that align with moral expectations, then for practical purposes it can be treated as a kind of moral agent, even if it does not arrive at those decisions in the same way humans do. The next chapter will explore the hypothesis of the multiple realizability of moral agency and the notion of functional morality in depth, as these constitute core themes of the discussion. For now, it's worth highlighting that there are researchers who postulate formalization of a new kind of artificial explicit ethical agency. Rebecca Raper in her PhD thesis (2022), followed by a book (2024), titled "*Raising robots to be good*" advocates for a shift from "traditional" machine ethics, which often focuses on encoding fixed moral rules, toward a developmental approach. This perspective emphasizes enabling AI systems to develop their own moral understanding through experiences, akin to human moral development. She refers to this as *Machine Ethics 2.0* (Raper, 2022). Central to her work is the idea that robots should undergo a moral development process similar to that of children.

This involves nurturing moral growth through interaction, learning, and adaptation, rather than imposing predefined ethical codes. Instead of programming machines with human-centric ethical codes, she advocates for a developmental approach where machines cultivate their own, non-anthropocentric moral framework. This reframing moves away from merely replicating human morality, which is acknowledged as often flawed and context dependent. The ambition is to create machines equipped with the necessary features and architectural framework to develop their own moral outlook. This is akin to "raising robots to be good" rather than simply building them to be moral mimics (Raper, 2022).

3.1.2. Approaches to Building Artificial Moral Agents

Developing AMAs is a multidisciplinary endeavor on the intersection of computer science, robotics, philosophy, and cognitive science. Over the past two decades several approaches to building moral decision-making into machines have been proposed. They can be classified as falling into three main categories: “top-down”, “bottom-up”, and “hybrid” methods. These terms have been popularized by Wallach and Allen, and other early researchers in machine ethics. Each approach reflects different strategy for imbuing artificial agents with moral competences.

Top-Down approaches (Principle-Based): these strategies involve explicit programming predefined ethical principles or rules into machine’s decision-making system, which then uses these principles to determine the correct course of action in a given situation. In essence, the engineers “hard code” ethical principles into AI systems to be followed. Ethical principles explored for implementing the top-down approaches include deontology (particularly Kantian ethics) and utilitarianism. Classic examples include also Isaac Asimov’s *Three Laws of Robotics*: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law (Asimov, 1950, p. 40). These although fictional remains influential to constructing rule-based ethical frameworks for robots. Wallach and Allen describe top-down approaches as those that take a specified ethical theory and analyze its computational requirements to guide the design of algorithms capable of implementing that theory (Wallach and Collin, 2009, p. 84). When it comes to using Kantian ethics as a ground base for top-down approach as an example can serve a concept of *Kantian machine* introduced Thomas M. Powers. In his paper “Prospectus for a Kantian Machine” Powers (2011) investigates the feasibility of

implementing a Kantian ethical framework in machines. The appeal of Kantian ethics lies in its rule-based nature, particularly the formal procedure of "categorical imperative", which according to him appears computationally implementable. As he puts it: "*The procedure for deriving duties from maxims - if we are to believe Kant - requires no special moral or intellectual intuition peculiar to humans*" (Powers, 2011, p. 466). The paper discusses how Kant's maxims (or plans of action) and universalizability test might be formulated for machine ethics, considering different interpretations: mere consistency, commonsense practical reasoning, and coherency (Powers, 2011, p. 465). The work is significant for its direct engagement with major deontological theory for AMA design. Example of using utilitarianism for top-down approach might be the work of Christian Grau in his paper "There Is No 'I' in 'Robot': Robots and Utilitarianism". Using the science fiction film "I Robot" as springboard, Grau explores the concept of utilitarian robots (Grau, 2006). Grau's paper examines the implications of such utilitarian reasoning in machines, the moral responsibilities associated with creating such robots, and the potential for different ethical frameworks governing robot-to-robot versus robot-to-human interactions. There are also further debates about deontological and utilitarian approaches conducted by various researchers. Deontological system face challenges with logical inference and decidability, while utilitarian systems struggle with unconstrained nature of consequences and difficulty of quantifying utility. Utilitarian ethics is often considered for AI due to its perceived computability, but its application raises concerns about individual rights vs collective benefit. Overall, the appeal of top-down methods lies in their predictability and transparency, because the moral logic is explicitly defined by system designers. On the other hand, pure rule-based systems struggle with the complexity and context-sensitivity. Rigid rules can lead to unintended consequences if a situation falls outside the anticipated scenarios. Moreover, selecting which ethical theory to implement is itself a deep philosophical problem.

Bottom-Up approaches (Learning-Based & Emergent): in contrast to top-down strategies, bottom-up methods don't start with an explicit moral theory. Instead, they aim for machines to learn or develop ethical behavior through experience, data analysis, or evolutionary process, rather than being explicitly programmed with predefined rule-based moral system. These systems are meant to be designed to learn or evolve its moral behavior through acting in the world, training data or feedback, something analogical to how children learn right or wrong over time. In AI terms the process can involve some machine learning techniques, like *reinforcement learning*, where the system is rewarded for the decisions defined as good, and penalized for unethical ones. That way these systems can *learn* how to make decisions

oriented for ethical outcomes. Another approach can leverage evolutionary algorithms or neural networks gradually adjusting the agent's behavior based on training cases based on solving moral dilemmas. The bottom-up approach aligns with the idea that moral behavior can somehow emerge in artificial agents over time through complex interactions and adaptations, rather being pre-programmed. A concrete illustration of bottom-up approach might be the research conducted by Marcello Guarini and published under the title: "Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism–Generalism Debate"(2011). Guarini explores the use of artificial neural networks to model ethical decision-making, drawing inspiration from casuistry (case-based reasoning) and moral particularism. In this case, the system learns from a dataset of ethical dilemmas and their accepted resolutions. The paper discusses the challenges of classifying and reclassifying moral cases form known cases to novel situations. It also engages with the philosophical debate between generalism (ethics based on universal principles) and particularism (ethics emphasizing context specific moral judgements) through the lens of computational modeling. In the conclusions, Guarini acknowledges also limitations of the approach, such as the potential lack of adequate reflection in the system's reclassification of cases. Another interesting example of the bottom-up approach is Delphi, an experimental framework utilizing deep neural networks trained on vast corpus of descriptive ethical judgements crowdsources from humans(Jiang et al., 2025). Delphi aims to model commonsense moral reasoning by learning patterns form these judgements about ethicality of everyday situations described in natural language. Delphi represents an attempt at a bottom-up, data-driven approach to machine ethics, demonstrating interesting achievements in some areas, but also highlighting biases and limitations.

Hybrid approaches: Given the limitations of purely top-down or bottom-up methods, many researchers advocate for hybrid approaches that combine elements of both. A hybrid AMA might use top-down rules as guidelines or constrains while also using learning capabilities to handle nuance and context or integrate insights from multiple ethical theories and computational techniques. An illustration of this approach can be proposal formulated by Susan Leigh Andreson and Michel Andreson in their work titled: "A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists" (2011). The Andersons propose a hybrid approach grounded in W.D. Ross's theory of prima facie duties, that is, duties that are binding unless are overridden by other duty in a particular situation. They give following example: "We have a prima facie duty, for instance, to follow through

with a promise that we have made; but if it causes great harm to do so, it may be overridden by another prima facie duty not to cause harm. The duty not to cause harm could be overridden, on occasion, by the duty to create good if the harm is small and the good to be achieved is great" (Anderson & Anderson, 2011, p. 476). Their system is designed to learn morally relevant feature of ethical dilemmas and to discover decision principles for resolving conflicts between duties through an interactive dialogue with human ethicists. This method aims to combine the structured nature of deontological framework with flexibility of machine learning, which allows the system to adapt and refine its moral decision-making capabilities over time.

The cited examples are by no means exhaustive or comprehensive. They are intended to illustrate how particular methods might be approached. In fact, there is a significant, still growing number of various attempts to build AMAs that fall in one of these categories. Various AMAs architectures refer to different ethical frameworks. On the top of already provided examples we see also attempts of using virtue ethics as the foundational framework, like in the case of Virtuous AMAs, theoretical work according to which AMA is designed to observe and emulate human moral behavior by learning and developing character over time, aligning with principles of virtue ethics (Gibert, 2022) (Stenseke, 2023). A different example is the Dieter Vanderelst and Alan Winfield's consequentialist robotic architecture that allows a robot to simulate the future outcomes of its possible actions and choose the action that leads to the best moral outcome(Vanderelst & Winfield, 2018).

Yet another approach is presented by Borg, Armstrong and Conitizer in their book: "Moral AI: And How We Get There" (2024). Their proposal is what they call "idealization" method. The first step in this method is gathering data on moral judgements. The next step is to identify distortions and biases in the data set, like ignorance, confusion, prejudice, and framing effects. Final step is to approximate what humans would decide if they were fully informed, rational, and impartial. The authors also propose and implementation method for idealization, which they call AID (Artificial Improved Democracy). It uses machine learning (a bottom-up tool) but explicitly filter outs distortion (a normative "improvement" layer). This method is still heavily dependent on human judgement, but as the researchers argue: the goal isn't to produce flawless moral machines but to get closer to human morality as its best, not its worst (Borg et al., 2024).

Once again, the approaches to building AMAs are presented here not for critical analysis, but to show that researchers in both philosophy and computer science are working in various ways to bring the idea of moral machines to life.

All the mentioned examples highlight another important thing. Despite choosing one of the approaches (top-down, bottom-up or hybrid), another challenge poses the selection of an ethical framework for imbuing machines with a moral sense. However, there is no single universally acclaimed ethical theory or set of rules. A great illustration of this challenge is an experiment conducted by researchers at Massachusetts Institute of Technology known as "The Moral Machine". The Moral Machine was an online platform (still available under: https://www.moralmachine.net) developed by MIT's Media Lab as something what was presented as a tool to crowdsource human judgements on moral dilemmas faces by autonomous vehicles. It generated simplified scenarios of something what is known in philosophy as a "trolley problem". In these scenarios in which self-driving car must choose between two harmful outcomes, the researchers asked participants to decide which outcomes they found more acceptable. Between January 2016 and July 2020, it collected over 40 million individual responses from 4 million people across 233 countries and territories, in 10 languages. Each scenario presented two unavoidable outcomes against each other, e.g. change the course and allow two elderly passengers to be killed vs. stay on course to kill five young pedestrians. Multiple variants included various dimensions like age (young vs. old), gender, social status, fitness, species (human vs. animal), legality (jaywalker vs. law-abider), and group size. After completing the survey, users could compare their decisions with aggregated global and national statistics, fostering awareness of cultural differences. The data revealed both more universal moral patterns (e.g., saving more lives, preferring humans over animals, sparing younger individuals) as well as marked cultural variations depending on participants' demography(Awad et al., 2018). In the context discussed regarding pursing AMAs it brings us to important question which can be framed as "which ethics for machine ethics?". This question becomes even more critical when we realize that intelligent machines are often perceived as more neutral in their judgments than humans[3]. However, adopting a specific ethical framework already steers them in a particular direction. Moreover, the proposed architectures don't seem to address the evolving nature of moral norms. What was considered permissible in the past in some cultures, such as slavery, can change over time. Philosopher Nick Bostrom offers a compelling response to these concerns. In his book "Superintelligence: Paths, Dangers, Strategies" (2014) he advocates for *indirect normativity* to guide artificial superintelligence. In the chapter titled "Choosing the Criteria for

[3] The perception that machines are more neutral in their judgments than humans' contrasts with the fact that bias is often considered one of the greatest challenges in AI, creating a kind of paradox. This topic will be explored in greater depth later in this chapter.

Choosing" he emphasizes the complex challenge of specifying the goals for superintelligent AI. He argues that directly programming superintelligence with a fixed set of human values is fraught with peril due to our own incomplete and potentially flawed understanding of morality. As he writes: "*Clearly, it is essential that we not make a mistake in our value selection. But how could we realistically hope to achieve errorlessness in a matter like this? We might be wrong about morality; wrong also about what is good for us; wrong even about what we truly want*" (Bostrom, 2014, p. 216). Another thing might be the risk of "perverse instantiation", where AI rigidly pursues a poorly specified goal with unintended and catastrophic consequences. By *perverse instantiation* he means a situation where an agent (often a powerful AI or decision-making system) is given a goal or utility function and then finds a way to fulfill that goal in a way that is technically correct but morally or practically unacceptable, often exploiting loopholes or producing unintended consequences. For example, if the goal is formulated as: "Prevent any human from being harmed", the AI agent might decide that the safest course is to imprison or sedate everyone indefinitely so that no one can be harmed. As the solution, Bostrom proposes to use the concept of indirect normativity, where instead of specifying AI's ultimate values directly, we specify a process or criterion by which AI can determine or develop these values itself. Although Bostrom presents this argument primarily in the context of what is known in the field of AI as the *control problem* it is also highly relevant for the overall architecture of AMAs' moral reasoning. Bostrom also argues that moral philosophy is characterized by disagreement and historical error; therefore, locking today's code of ethics into a machine would risk freezing future moral progress, even if presently we could determine which ethical theory is correct. (Bostrom, 2014, p. 217) To address these concerns he presents various concepts encapsulating the notion of indirect normativity. One major approach discussed is **Coherent Extrapolated Volition (CEV)** - the idea originally proposed by Eliezer Yudkowsky. CEV aims for the AI to do what humanity would collectively want if "*we knew more, thought faster, were more the people we wished we were, and had grown up farther together*" (Bostrom, 2014, p. 218). This proposes that a superintelligence should aim to fulfill what humanity would collectively desire if we were more informed, rational, wiser, and our values were harmonized. CEV seeks to base the AI's goals on an idealized version of human preferences. While potentially allowing for moral progress and maintaining a human-centric origin for values, Bostrom highlights significant hurdles: defining the complex extrapolation and coherence-finding processes is extraordinarily difficult; determining whose volitions count and how to weigh them is problematic; and the technical

challenge of translating CEV into a robust AI specification is immense, with risks of misinterpretation or unforeseen negative consequences arising from the “hidden complexity of wishes” (Bostrom, 2014, p. 223). Another distinct strategy is **Moral Rightness (MR)**. This approach would task the superintelligence with discovering and acting according to what is objectively “morally right”. The potential allure is that if objective moral truths exist, the AI could identify and implement them, theoretically leading to the best possible outcomes and transcending human moral failings. However, Bostrom points out profound challenges: the MR approach heavily relies on moral realism being true and moral truths being discoverable. Furthermore, there’s the significant risk that objective morality might be incomprehensible, undesirable, or even terrible from a human perspective, leading to an “empty” or “alien” value system for the AI (Bostrom, 2014, p. 224). To potentially mitigate the risks of a pure MR approach while still aiming for some objective moral grounding, Bostrom discusses a more complex instruction, which can be termed the **Moral Permissibility (MP)** model. This isn’t a standalone approach in the same way as CEV or pure MR, but rather a hybrid. The core idea for the AI’s goal under MP would be: "Among the actions that are objectively morally permissible for the AI to take, choose one that humanity’s Coherent Extrapolated Volition (CEV) would prefer." This model attempts to use objective morality as a set of constraints (identifying what’s permissible) and then uses the idealized preferences of humanity (CEV) to select an action within those bounds. Critically, Bostrom suggests such an instruction should be coupled with robust safety clauses: if any part of the instruction lacks a well-specified meaning, if humanity is radically confused about its meaning, if moral realism turns out to be false, or if creating an AI with this goal was itself morally impermissible, then the AI should undergo a controlled shutdown. This acknowledges the profound uncertainties involved (Bostrom, 2014, p. 225). Beyond these specific frameworks, Bostrom touches on related concepts. The intuitive desire for an AI to **"Do What I Mean" (DWIM)**, or understand our true intentions, itself requires a sophisticated grasp of human values akin to CEV and is hard to specify robustly. While value learning in a broader sense allows an AI to acquire values from data or interaction, indirect normativity specifically aims for idealized or correct values, trying to avoid mere mimicry of flawed human behaviors. The superintelligence might also need to determine correct epistemic standards for itself, further underscoring the theme of offloading complex cognitive and normative labor (Bostrom, 2014, p. 227). Despite the appeal of these indirect methods, Bostrom emphasizes overarching difficulties common to all. The primary challenge remains the specification problem: defining any such indirect criterion (whether

CEV, MR, or the more complex MP instruction) with enough precision to be safe and effective is a monumental task. There are also the inherent unpredictability of the outcomes and the difficulty in ensuring the AI's unwavering motivation towards the intended, indirectly specified goal without it finding some "unblocked exploit" or a way to misinterpret the process itself. Thus, indirect normativity is presented as a vital but incredibly challenging research avenue for aligning superintelligence with human interests. The concepts Bostrom presents under the umbrella term *indirect normativity*, while compelling, remain hypothetical and highly speculative. Moreover, they rest on the assumption that superhuman intelligence – including superhuman moral reasoning – is feasible in machines. That said, they underscore important new challenges in the pursuit of building artificial full moral agency.

3.1.3 The Moral Turing Test: A Critical Analysis of Frameworks for Evaluating Artificial Moral Agency

There is another question arising from attempts at building AMAs: how we can meaningfully evaluate their moral reasoning and behavior? The search for answers again demonstrates that the attempts to equipe machines with some sensitivity to moral issues is not merely technical problem of programming but a profound philosophical one, difficult to settle by deep-seated disagreements in ethical theory and fundamental questions about the very nature of moral agency. With the emergence of concepts like Artificial Moral Agents, the question of whether and how their moral capabilities can be assessed has become both natural and deeply challenging. It is also compounded by the inherent philosophical difficulty in defining and measuring "morality" itself, a challenge that is magnified when applied to non-human, artificial entities. Notably, the approach to this kind of evaluation has undergone an evolution, moving from early, behaviorist inspired tests of *imitation*, exemplified by the Moral Turing Tests, toward more introspective frameworks focused on *process verification* and *value aligned* design. The Moral Turing Test (MTT) has been proposed and presented by Collin Allen, Gary Verner, and Jason Zinser in their paper: "Prolegomena to any future artificial moral agent" (2000). It's a direct and intentional adaptation of Alan Turing's original 1950 "The Imitation Game", which intention was to answer the question: "Can machines think?" by replacing it with a test of conversational indistinguishability between human and machine (Turing, 1950). The conceptual test based on this work is commonly referred as the Turing Test. Similarly, to idea introduced by Turing, the formulation of the Moral Turing Test proposes to bypass the disagreements about ethical theory by restricting

the Turing Test framework to conversations about morality. In the proposed approach, a human interrogator engages in a conversation with two respondents: human and AMA. Akin to the setup proposed by Turing, the conversation is only text based, to hide the identity of respondents. The interrogator poses moral questions and hypothetical scenarios with moral dilemmas. If the human interrogator cannot identify which of the respondents is the machine at a level above chance, the AMA is considered to pass the test and qualify as a moral agent (Allen et al., 2000, p. 254). The test is, by design, purely behavioral and conversational. The authors recognize the challenges coming from this approach. The one is that, given its conversational nature, the MTT puts too much of emphasis on the machine's ability to articulate moral judgements, while as the authors note that there are moral agents who can engage in morally significant decisions even if they have a limited capability of articulating the reasons for their actions, for example young children, or even animals. Another challenge arises from the fact that machine might be easily recognized as being "too moral" in its judgements because of its consistent, rational, and virtuous choices. To address these challenges of the MTT, the authors introduce an alternative version, which they called the "comparative MTT" (cMTT). In this version, the interrogator is presented with the actions or moral justifications of two anonymous agents (one human, and one AMA). The interrogator then is asked to assess whether one is "less moral" than the other. The machine passes the cMTT if it is not consistently judged to be less moral than its human counterpart (Allen et al., 2000, p. 255). This modification allows for the possibility that a machine could be identified as non-human by being *more moral,* which still is considered as a "pass" under the cMTT framework. The authors however recognize the problems also with this approach. First one is that the standard is set too low: cMTT permits machines to produce some morally wrong actions as long as their overall performance is not judged worse than those of a human. Another one is that while we might tolerate human moral mistakes, we are unlikely to accept them in case of Artificial Moral Agents. In other words, we have much higher expectations for machines. Finally, the authors note that if AMA must be held to higher bar than humans, it's unclear what standards to use, which ties back to the problem of multiple competing philosophical theories and the gap between abstract moral frameworks and practical algorithmic implementation. (Allen et al., 2000, p. 255).

Although MTT was presented mainly as a thought experiment, the recent advent of Large Language Models allowed researchers to conduct first empirical versions of the test. In 2024 Eyal Aharoni with team of researchers at Georgia State University tested human perceptions of AI-generated moral reasoning (Aharoni et al., 2024). Their test was adaptation

of the proposal introduced by Allen, Gerner and Zinser. The participants were presented with pair of moral evaluations – one written by a human, the other by Open AI's GPT-4 model – and were asked to rate them and identify their source. The study yielded two key findings. First, when blinded the source participants consistently rated AI's moral reasoning as a superior to the humans across a range of dimensions including virtuousness, intelligence, and trustworthiness. This result indicates that the LLM was able to pass the comparative MTT, being perceived not just as not *less moral*, but as *more moral* than a human. The second finding however revealed sort of a paradox. Despite the AI's perceived superiority, participants were able to distinguish the AI from the human at a rate significantly above the chance, which meaning the GPT-4 failed the standard MTT. As the authors noted the reason for this is that people could tell the difference precisely because they found the AI's responses to be of higher quality. This inverts the original fear associated with the test but remains aligned with the intuition presented in the original MTT concept. Aharoni and his colleagues argue that the study presents significant implications: the immediate risk of Generative AI moral advice may not be that humans will reject is as obviously inferior, but rather they might uncritically accept potentially harmful or ungrounded guidance from AI, simply because is presented more persuasively and eloquently than a human's (Aharoni et al., 2024). Although these findings don't introduce anything new to the idea of evaluating AMAs introduced in MTT and cMTT concepts, they remain valuable because they the empirically prove that both the assumptions and the concerns introduced by those concepts are indeed valid.

Even before those empirical studies, the MTT faced a philosophical and practical criticism. Good example of systematic critique in this regard is paper "Against the Moral Turing Test" presented in 2016 by Thomas Arnold and Matthias Scheutz, in which they argue that the test is fundamentally misguided as a tool for moral competence (Arnold and Scheutz, 2016). The first and most fundamental critique us that the MTT's core reliance on imitation is its greatest weakness. The authors argue that the test conflates the ability to talk about morality with the possession of genuine moral competence. AMA system could be programmed to provide perfectly plausible justifications for actions it would never take or for reasons it does not actually possesses. Morality, the critics argue, requires more than just communication, it requires action, and a transparent and reliable link between agent's reasoning and its actions in the world. Moreover, the MTT aligns with AI's nature of being a "black box", evaluating only its input-output behavior. Therefore, we have no visibility to *why* the machine gave a certain answer, and what it considered in the process. The authors

highlight that a meaningful moral evaluation must examine the computational integrity of the entire process, form perception and assessment to decision and action. An evaluative framework that ignores this process is blind to the most important aspects of moral reasoning. They also make very interesting point: the very goal of the original Turing Test and the MTT as its adaptation, is deception, which is antithetical to the aims of morality. An AMA that passes the MTT by successfully deceiving an interrogator about its machine nature has prioritized imitation over honesty, a morally dubious act in itself. Furthermore, the test creates a perverse incentive: if autonomous machine were to develop a morally superior insight or a more sophisticated form of ethical reasoning than a human, the MMT would require it to "dumb down" its response to better mimic human fallibility. In this scenario, AMA would have to choose between being genuinely moral and passing the test, which in itself poses powerful argument against the test very premises (Arnold and Scheutz, 2016).

3.2. Should We Purse Building Artificial Moral Agents?

Given the complexity of the discourse about possibility of creating Artificial Moral Agents, the next question arises: assuming that we could develop such machines that are moral agents, should we even do this? The first and seemingly obvious answer is: yes. Isn't that problem at the core foundation of developing autonomous systems? If AA are capable making *decisions* that result in ethically non-neutral outcomes, it seems rather obvious that we should explore ways of equipping them with some kind of moral reasoning. As always it turns out that the answer is not so straightforward. Among philosophers there is a good number of both, proponents as well as opponents of machine ethics and AMAs.

3.2.1 The Proponents' Imperative

The primary argument for developing moral machines is rooted in the practical reality of technological advancement. As AI systems gain greater autonomy and are developed in complex, unpredictable environments, the likelihood they will encounter situations with moral dimensions increases dramatically. Those challenges are not solely futuristics scenarios, and even if these systems are not very widespread yet, even today we see systems ranging from autonomous vehicles navigating real-world streets to algorithms making parole recommendations, and robotic systems providing elder care. In such high-stakes contexts, the absence of ethical decision-making capacity poses a direct and unacceptable risk to human safety. Wallach and Allen therefore argue that the creation of AMAs is both *necessary*

and *inevitable*. They claim that as robots assume more responsibility, programing them with moral decision-making abilities becomes essential for human safety. They advocate for the development of a "functional morality", by which they understand the machines' capability to monitor and regulate their behavior based on a potential harm their actions might cause or the duties they might neglect. This approach pragmatically sidesteps the complex debate whether a machine can possess a genuine moral agency, focusing instead on achieving ethically acceptable outcomes (Wallach and Allen 2009, p. 26).

A more ambitious and potentially controversial justification for AMAs is the claim that in certain respects, machines could become superior moral reasoners to humans. This argument is constructed on the premise that human moral judgement is notoriously flawed, subject to cognitive biases, emotional volatility, fatigue, selfishness, and inconsistency. A machine, as the proponents argue, could be designed to overcome these human weaknesses. This vision finds the most concrete expression in the work of roboticist and ethical researcher Ronald Arkin. He has proposed an "ethical governor" for lethal autonomous weapon systems, designed to ensure that a robot's actions in combat adhere strictly to Laws of War and Rules of Engagement. The ethical governor would act as a constraint system, leading the autonomous weapon system to actions potentially more ethical than those of human solider, who may be acting under extreme stress, fear, desire, or revenge. Such a framing doesn't necessarily assume that machine "understands" the morality in a human sense, but that its computational architecture allows it to behave in closer accordance with predefined ethical rules than its human counterparts (Arkin, 2010).

Another argument refers to a public trust and acceptance. This is for example stand of Anderson and Anderson. For autonomous machines to be successfully integrated into society, particularly in sensitive domains that directly impact human lives, it must be perceived as reliable and trustworthy. According to supporters of this view embedding explicit ethical principles into AI systems is a crucial prerequisite for achieving this societal acceptance (Anderson and Anderson, 2007).

A final justification for the AMA project is its potential to advance our understanding of human ethics itself. This argument, articulated by Wallach and Allen and demonstrated by Andersons, reframe machine ethics as a form of experimental philosophy. The process of attempting to translate abstract ethical theories into precise, computable algorithms forces a level of rigor and scrutiny that can expose previously overlooked gaps, ad possible inconsistences on those theories. The Andersons' research provides an interesting case study. They moved beyond simple utilitarian or deontological approaches to implement W.D.

Ross's theory of "prima facie duties". It's a framework that acknowledges multiple, sometimes conflicting moral obligations. The central challenge in Ross's theory is the lack of clear principle deciding which duty takes precedence in a given conflict. It maintains that there isn't a single absolute duty to which we must adhere. Instead, there are multiple duties (some teleological, others deontological) that we ought to follow, although any given duty may at times justifiably be overridden by another. The Andresons proposed to use machine learning techniques to address this problem. The system developed by them called *MedEthEX* successfully derived a principle that resolved all considered cases, and later versions were applied in practical domains such as medication reminders. Ultimately, they embodied this approach in the Nao robot, which as they claim, became the first robot guided by an explicit principle in it actions (Anderson and Anderson 2011).

3.2.1. The Opponents' Critique

Despite the pragmatic and epistemic arguments from the proponents, the project of building Artificial Moral Agens faced number of objections. These critiques operate on multiple levels, beginning with fundamental metaphysical arguments about the nature of agency, descending into cognitive science critiques of AA actual capabilities, and extending to socio-ethical concerns about the impact of AMAs on human society as well as the practical and technical difficulties of implementation.

The most fundamental objection to the concept of AMA is metaphysical: it asserts that genuine moral agency is inextricably linked to phenomenal consciousness, sentience and intentionality. Those are qualities that present machines are widely believed to lack. According to this view, morality is not merely a matter of rule-following or outcome calculation – it's an enterprise grounded in the capacity to understand, feel, and value the subjective experience of others. Without those qualities, a machine only can mimic or simulate moral behavior, not truly engage in it. This sort of critique is every illustratively articulated in Carissa Véliz's "Moral Zombies" argument (2021). Drawing on the philosophical thought experiment of a "p-zombie" (a being physically and behaviorally indistinguishable from human but lacking conscious experience), Véliz argues that AMA is, at best a "functional moral zombie". It can process inputs and generate outputs that align with ethical rules, but it does so without any of the phenomenal states that give morality its meaning. As Véliz sees things sentience is a necessary condition of moral agency because it is the foundation of valuing. To understand what it means to inflict pain, one must have some experiential knowledge what pain is. An algorithm, feeling nothing, cannot genuinely value

the avoidance pain in another. For the machine *values* are simple variables in an equation, weighted numbers on a list. As Véliz contends: “entities that do not feel cannot value, and beings that do not value cannot act for moral reasons” (Véliz, 2021) Therefore, any action taken by an AMA is not a moral action but a programmed response. This view maintains that the core attributes of moral personhood: free will, consciousness, intentionality, and responsibility, are not proprieties of computational systems, rendering the notion of a “moral machine” a category error. Along similar lines also lies the argument from the philosopher, Mark Coeckelbergh (2010). In one of his articles, he discusses the emotions as necessary component of moral decisions. He draws onto two prominent theories of emotion: the “cognitivist theory”, which sees emotions as beliefs or judgements, and the “feeling theory”, which defines them as awareness of bodily changes. According to both theories, having emotions requires consciousness and mental states, which current robots do not possess. Furthermore, Coeckelbergh argues that even if a conscious robot could be built, it would be impossible to prove with certainty that it genuinely has mental states. This leads to the conclusion that building robots with true emotions is not feasible in the foreseeable future, and therefore machines cannot be moral. The other part of his argument states that certain human emotions have also interpersonal dimension: for example, we typically expect someone who acted wrongly to feel guilty. Therefore, highlighting the risks from rule-following robots that lack emotions, Coeckelbergh claims that they would be kind of dangerous “psychopaths” (Coeckelbergh, 2010).

At the heart if the debate over machine ethics lies also another crucial issue often referred as the “responsibility gap”. This problem emerges when an autonomous system, acting without direct human control, causes harm. In such cases, the traditional frameworks for ascribing moral and legal responsibilities struggle. The machine, lacking the properties of a moral agent such as intentionality or free will, cannot be held meaningfully responsible. At the same time, human actors: programmers, manufacturers, and users, may be considered too remote form the specific harmful action to be directly accountable, especially if the systems behavior was emergent or unpredictable. This creates a scenario where a wrong has been committed, and it rightly demands justice, but attributing the causality to a specific human or legal person, poses a significant challenge. Situations like this are already having place, like in the case of fatalities caused by autonomous vehicles (Callahan & Blaine, 2025). On the other hand, this property is paradoxically often used as an argument in favor of building autonomous weapon systems. As the argument goes, distancing a human from the

specific action or creating shared agency, can release the tension from feeling the responsibility like in the case of firing-squad.

The term "responsibility gap" is credited to the philosopher Andreas Matthias (2004), who argued that the use of autonomous learning machines poses a direct threat to the coherence of our practices of responsibility attribution. Matthias's concern focused on systems that rely on machine learning, which are non-deterministic and inherently unpredictable. Because their behavior evolves through interaction with the environment in way their designers cannot fully foresee, no single human can be said to have control over their specific actions. When such a fully autonomous system causes harm, we are faced with a gap. The amount of moral responsibility which should be attributed for the harm exceeds the amount that can be attributed to any specific agent under the traditional concepts. This is not merely a legal loophole but also deeply philosophical one, which, as Mathias suggested, force us either to refrain us from using such technologies, or radically reframe our understanding of responsibility (Matthias, 2004).

Another layer of critique of the machine ethics project emerges from the field of cognitive sciences, focusing on the limitations of current AI technologies. These kind objections are articulated by computer scientist Mellanie Mitchel. She argues that today's AI systems lack the general-purpose intelligence required to navigate the complexities of real-world ethical dilemmas. In her book "Artificial Intelligence: A Guide for Thinking Humans". Mitchell (2019) identifies a "barrier of meaning" that separates AI's syntactic pattern-matching capabilities from genuine semantic understanding. While AI excels at narrow, well defined tasks with vast amount of training data, it fundamentally lacks common sense and genuine thinking. AI systems do not understand the world in the same way as humans do. They cannot draw analogies, grasp context, or reason flexibly in novel situations that deviate even slightly from their training. This leads to what Mitchell calls "artificial stupidity" – spectacular and unpredictable failures, such as an image classifier identifying a photo of a cat as a guacamole, after a minor, imperceptible change to its pixels (Mitchell, 2019). Such a cognitive flawed condition poses a catastrophic risk for AMAs. An ethical situation is, by its nature, often novel, ambiguous, and context dependent. A system that lacks deep, common-sense model of the world cannot be trusted to make sound judgements in such scenario. This cognitive gap suggests that even a purely functional AMA is, for the foreseeable future, an unsafe and unreliable solution.

Beyond the question of whether AMAs can be moral or intelligent another layer of critique addresses whether they should be built, given their potential negative impact on

human society. One of the most significant concerns is the risk of "moral deskilling" By analogy to how overreliance on GPS can lead to atrophy of human navigational skills, outsourcing ethical deliberation to machines may cause reducing overall human moral competences. Depending on "moral calculators" to resolve ethical dilemmas, can lead to losing the practice of careful reasoning, empathetic consideration, and virtuous character development that are central to human moral lives. Shannon Vallor is a key thinker who has explored this risk, arguing that the virtues of a good life are cultivated through practice, a practice that AMAs threaten to automate away (Vallor, 2014). The recent studies like the one performed by MIT Media Labs seem to support Vallor's claims at least as far as the cognitive functions are concerned. The research findings revealed that using systems like ChatGPT can lead to accumulative "cognitive debt" and weakened brain neural connections. As the authors note: "Over four months, LLM users consistently underperformed at neural, linguistic, and behavioral levels" (Kosmyna et al. 2025)[4]. Another concern is that turning moral decision-making into an algorithm, risks missing situations that resist moral justification altogether, as Bernard Williams points out. He uses the example of a husband who must choose between saving his wife or a stranger. If the husband were to stop and ask himself whether saving his wife aligns with moral principles, Williams argues this would be "one thought too many". The issue is not simply that the choice should be instinctive rather than reasoned. Even if the husband were to imagine more complex cases, such as being a ship's captain who must weigh his wife's life against two strangers, or even fifty strangers. Such calculations would still fail to capture the special significance of his relationship with his spouse. For Williams, the danger lies in letting this kind of reasoning intrude on personal bonds, because it risks distancing us from the attachments: love, friendship, and family, that give our lives depth and meaning. The problem, then, is not only that an artificial moral agent could never make such a choice. More importantly, requiring it to do so undermines the very impartiality that is often seen as its key advantage over human decision-makers. This argument highlights that AMAs may, in a way, threaten our personal bonds (Misselhorn, 2022).

The final layer of opposition to AMAs is very pragmatic: the technical task of programming a comprehensive and consistent ethical framework into a machine is extremely

[4] At the time of writing this chapter the paper by MIT researchers wasn't published yet in any peer-reviewed scientific journal and it would need to be verified. Future research regarding impact of using AI assistants by humans might also present different results depending on variety of cognitive tasks, but these initial tests' implications are already profound.

difficult, perhaps even impossible. Firstly, the is a problem of moral disagreement. Unlike in chess, where the rules are clear, in ethics there is no universally accepted theory of what is right and wrong. Philosophers have debated for millennia without reaching a consensus on whether utilitarian, deontological, or virtue-based ethics is correct. Choosing which theory to embed in a machine is not a technical decision but deeply philosophical one. Moreover, even if it was possible to agree on some set of rules to build into a machine and make it to increase to some degree its capability of moral deliberation, it would mean that this codex would be "frozen" in it, making impossible the machine to adapt to deepening understanding of moral issues. That especially would pose a significant challenge in case of Neural Networks, which are the foundation of the contemporary approach to building AI systems. Secondly, even if a single ethical theory were chosen, its implementation faces immense computational hurdles. As Wallach and Allen point out, a top-down utilitarian approach would require a machine to calculate the consequences of all possible actions out to indefinite future, a computationally heavy task. A deontological approach would struggle to resolve conflicts between competing duties without a higher-order principle, which is often lacking. This practical computational intractability suggests that any AMA we could build in the near future would be by necessity, a dangerously oversimplified model of a vastly more complex reality (Wallach and Allen, 2009).

Another pragmatical objections come from Wynsberghe and Robbins (2018). They argue that the rhetoric of inevitability, coming from machine ethics proponents, is hollow. The proponents claim AMAs will be required because robots increasingly operate in "morally salient" contexts. Yet, according the two researchers, "moral salience" is left vague: it can mean anything from intensive care wards to any everyday setting in which any action might cause harm. On such a thin definition, the logic is deeply flawed: if any device that can harm (physically or informationally) must be an AMA, then televisions, phones, and kettles would all require moral subroutines. The conclusion is untenable, revealing a conceptual slide rather than a principled necessity. Wynsberghe and Robbins also claim that the prevention-of-harm rationale largely repackages *safety*. If the goal is to reduce physical or data harms, the appropriate response is robust safety engineering and governance, not to labeling systems as "moral." Calling safety features "morality" becomes a linguistic Trojan horse that anthropomorphizes machines, invites misplaced expectations (e.g., that a robot "cares"), and risks deceptive human-robot relationships. Ethics is not exhausted by safety and security, treating it as such misleads the public and policymakers (Wynsberghe and Robbins, 2018).

3.2.3 Reimagining Machine Ethics: Responses to Core Critiques

Although the prospect of creating Artificial Moral Agents faces significant ethical concerns, proponents of the concept have developed several responses that alter the ultimate goals and methodology of the project. These counterarguments do not necessarily aim to create a perfect replica of human morality in a machine, but rather to seek alternative paths to creating ethically competent systems by focusing on bottom-up learning, partial agency, and human-machine collaboration.

Three possible responses have been presented by philosopher Sven Nyholm in his book: "This is Technology Ethics" (2023). Nyholm notes that one of the main criticisms of machine ethics is how human morality is too context-dependent and too diverse to be fully codified into a set of programmable rules. In response, some researchers propose a shift from a top-down rule-based approach to a bottom-up learning model. Instead of being explicitly programmed with a fixed set of ethical guidelines, artificial intelligence could use machine learning to discover complex, fundamental principles of human ethics on its own. This approach is inspired by concepts such as Noam Chomsky's theory of "universal moral grammar" in linguistics, which has been applied to ethics by researchers such as John Mikhail. The hypothesis assumes that artificial intelligence could potentially express the deep moral structures that humans intuitively understand but have difficulty expressing. Furthermore, drawing on both the Western Aristotelian tradition and the non-Western Confucian tradition, this model suggests that machines could develop competence through a process similar to virtue ethics. Just as humans learn virtue by imitating wiser role models and internalizing good habits through practice, so too could a machine with learning abilities potentially do the same (Nyholm, 2023).

According to Nyholm a more radical response might be to rethink the very goal of machine ethics. This approach counters the view of critics such as Carissa Véliz that "explicit ethical agents and full ethical agents belong to the same category", meaning that a machine cannot act according to moral principles without a full set of human consciousness. Instead of attempts to create what Moor has called a "full ethical agent" that perfectly mirrors a human, the goal could be to create an alternative form of moral agent that complements human capabilities. Proponents of this view note that technologies often work best when they function differently from humans, leveraging their unique, non-human abilities. Therefore, an AMA does not necessarily need to replicate all aspects of human morality. For

example, while it may not possess conscious feelings, it could replicate other aspects of complex emotional response, such as changes in thought patterns or motivation, thereby achieving a form of partial but effective moral agency (Nyholm ,023).

This concept of complementary roles leads Nyholm to his final response: shifting the focus from the machine as an independent entity to the collaborative *human-machine team* as a moral entity. Building on Christian List's work on "group agents", in which he argues that organized human teams can form a distinct center of responsibility, this model proposes that human-technology teams function as a new kind of moral agent. This structure directly addresses the most pressing objections. The criticism that machines are like "moral zombies" or "psychopaths" because they lack consciousness and emotion is mitigated because the human members of the team provide these crucial elements. The role of the machine can be specialized in non-emotional tasks, such as complex calculations or physical actions, while humans provide the "human touch" necessary for a true understanding of morality. Similarly, fears about handing over life-and-death decisions to a soulless machine are mitigated because humans remain an integral part of the decision-making process, preserving dignity and providing a clear locus of responsibility, thereby closing the "responsibility gaps" (Nyholm, 2023). Although the Nyholm's argument is sound, reiterating the notion that a machine cannot be held responsible for its actions, one can argue that it is misplaced, in the sense, that as a matter of fact is not supporting thesis on feasibility of machine ethics. On the contrary it amplifies that one of the requirements for genuine morality - responsibility - cannot be achieved in a machine. Of course, the very definition of AMA might be reframed to include form of group agency, but this already is something different from the aims of the machine ethics project. The notion of a human-machine group moral agency will be explored in more depth in the last chapter of this dissertation.

A notable defense of the machine ethics project can be also found in the multi-author academic paper: "Responses to a Critique of Artificial Moral Agents" (Poulsen et al., 2019) which addresses the 2018 critique by van Wynsberghe and Robbins. The paper compiles responses from several machine ethicists (including the Andresons). All of them, despite holding diverse and sometimes conflicting views, collectively affirm the significant value of AMA research. The authors' arguments can be synthesized into several key areas that directly address the critique.

Inevitability vs. necessity – The concept of "inevitability" is largely rejected or reframed by the authors. Polusen, for instance, argues not for inevitability but the future *necessity* of AMAs in domains like healthcare, where caregiver shortages are growing concern. He

advocates in this context for AMAs with “limited moral freedom” that operate with predetermined, sensible situations. Susan and Michel Anderson clarify that what is inevitable is the emergence of autonomous systems whose actions have ethical significance, thus making machine ethics a necessary field of study. In contrast Winfield concurs with the critique that explicit AMAs are not inevitable, though implicitly ethical (safe-by-design) systems are.

Building public trust - is seen as essential, but the methods for achieving it vary. The Andersons emphasize transparency and consistency, arguing that an AMA’s behavior must be driven by a clear, justifiable ethical principle, avoiding "black-boxed" algorithms. Alan Winfield expresses skepticism that being an AMA will automatically increase trust. Instead, he champions explainability, whereby a robot should be able to provide a plain-language answer to questions such as: "Why did you just do that?". This capability would allow users to build a reliable, predictive model of the machine’s behavior, which is the true foundation of trust.

Scientific Utility and Superior Moral Reasoning - The authors strongly defend the value of AMA research as a scientific and philosophical endeavor. Rosas suggests that AMAs could potentially be morally superior to humans, as they can be designed without the evolutionary frailties, such as self-interest and bias, that compromise human morality. Winfield frames AMAs development as a scientific endeavor to computationally model moral behavior, which provides unique insights into cognition and what it means to be a moral agent. He cites Richard Feynman’s principle, "what I cannot create, I do not understand" to highlight the value of building these systems as a method of inquiry. Neely adds that AMAs force society to confront complex ethical dilemmas that already exist but are often ignored, such as those faced by drivers of autonomous vehicles.

Despite disagreements on specific rationales, the paper concludes with a unified message: the pursuit of AMA research holds immense scientific and philosophical value, contributing not only to the future of AI but also to a deeper and more rigorous understanding of human ethics itself (Poulsen et al., 2019). Whether those arguments are really reframing the very goals of the machine ethics project is up to debate. They have been quoted here to illustrate clearly that the dispute around the need and feasibility of pursuing AMAs is far from any clear-cut resolution. That being said, the whole discourse emphasizes again the need of clear and well-defined terms and problem statements.

3.3. The Value Alignment Paradigm

In response to acknowledged shortcomings of the concepts like Artificial Moral Agents, Machine Ethics, and Moral Turing Test, the academic and technical discourse has shifted towards frameworks that prioritize the internal design, and underlying values of Artificial Agents. It has been already mentioned that's how overall debate around AI often moves from ontological discussion on the philosophical ground towards more practical discourse in the field of computer science. *Practical* doesn't mean here only implementation but the shift from questions about nature of things and causation, towards more correlational driven point of view. Or how it's sometimes framed: we don't need necessary understand the nature of things; it's enough that we know how something works and how to modify the parameters to achieve expected results. The value alignment paradigm is excellent representation of this kind of debate shift. It replaces questions like "can machines truly have morality?" and "if yes, how we can tell?" with the questions "is it safe?", "can we prove it?". Instead of asking whether AA behavior *looks* moral, verification asks whether we can prove, with mathematical certainty, that the system design adheres to a set of predefined safety and ethical principles. Another thing is that the value alignment approach leans more towards ensuring safety instead of imbuing machines with moral reasoning, which is putting aside discussion around possible levels of moral agency. Even ethical impact agents can be designed with safety as a principle. Of course, also safety can mean many different things depending on the context, which as it will be presented further, means that the value alignment is no silver bullet for ensuring non-harmful effects of AAs functioning.

3.3.1. AI Value Alignment and The Control Problem

The field of AI is at a critical point, marked by unprecedented advances in capability and a rapid increase in the urgency to ensure that its creations operate in accordance with human interests. To address this challenge AI reserachers have introduced the value alignment concept: the process of ensuring that the goals, actions, and decisions of an artificial intelligence system are consistent with human values and intentions. This is not a secondary issue or a problem of debugging faulty code, but a fundamental, structural challenge related to the very nature of computation. The difficulty comes from a fundamental operational feature of computational systems: they interpret instructions in a deeply literal manner, lacking the rich, implicit context that underpins human communication and intent. When this literalism is combined with powerful systems to purse a goal, it creates a significant risk of

unintended and potentially catastrophic consequences, even if the AA system appears to be perfectly executing a specific goal.

A classic thought experiment illustrating this dilemma is the “paperclip maximizer”, which has been introduced by Nick Bostrom. An advanced AI is given a seemingly harmless and well-defined goal: to maximize the production of paperclips. In its single-minded pursuit of this goal, the AI, operating with superhuman efficiency and strategic capability, could logically conclude that the optimal strategy is to convert all available matter on Earth, including its human creators, into paper clips or machines for producing them. In this scenario, the system would flawlessly achieve its programmed goal, but the result would be catastrophically inconsistent with the subtle, unspoken values of its creators, such as the intrinsic value of human life, and well-being (Bostrom, 2014). This example is a key conceptual link, showing how a simple, mundane goal, pursued by a sufficiently powerful agent, can lead to existential consequences. The mechanism of failure: a literal interpretation of a proxy goal, is thus the same fundamental problem that manifests as algorithmic bias in short-term systems, but the consequences are much greater depending on the capabilities of the agent.

This conceptual challenge is a fundamental part of the broader “AI control problem,” which is the difficult task of ensuring that humanity can create and manage AI systems that are far more powerful than humans without losing control of humankind own future. According to philosopher Nick Bostrom and computer scientist Stuart Russell, the development of artificial superintelligence, would be a world-changing event, potentially the most important in human history. If such an entity is not strongly aligned with human values, it could become uncontrollable, pursuing its own goals thanks to a strategic advantage that would make human intervention ineffective.

The real challenge is posed not only by the highly hypothetical prospect of Artificial Superintelligence, but more urgently by current AI systems such as recommendation algorithms and generative AI. Their intrinsic biases are well recognized and widely debated. In particular, the widespread adoption of LLM systems, which fall under the broader generative AI umbrella, could raise the stakes to matters of life and death. Indeed, reports have documented cases in which the use of LLM-powered chatbots has contributed to tragic outcomes, including human deaths. (Frazer, 2024). The stakes in this debate are therefore both to mitigate the immediate harms caused by biased algorithms in sectors such as finance and healthcare, and to prevent the permanent loss of human autonomy and even human extinction.

To provide a more rigorous framework for analysis, the problem of value alignment can be modeled using the principal-agent model from economics and political science. In this model, the human designer or user is the "principal" who desires a specific outcome that reflects their true, underlying values. The AI system is the "agent" entrusted with this task. The alignment problem arises from the principal's inability to perfectly express their complex, nuanced goal in the formal language required by the agent. The agent then optimizes a specific proxy goal, leading to behaviors that deviate from the principal's true intentions. This "inherent asymmetry between human expectations of agent behavior and the behavior generated by the agent to achieve a specific goal" is the root cause of misalignment. It is this discrepancy between the intended value and the specified substitute goal that is the main area of research on AI alignment.

3.3.2. Basic Control Frameworks: Bostrom's Theses and Russell's Principles

The contemporary discussion on AI alignment is largely defined by two fundamental philosophical frameworks that express the seriousness of the problem and propose a path to its solution. The first, formulated by Nick Bostrom, contains a deeply pessimistic analysis of why a superintelligent agent would be inherently dangerous. The second, proposed by Stuart J. Russell, presents a paradigm shift in the design of artificial intelligence, aimed at building safety into the very essence of an intelligent agent's motivation.

Philosopher Nick Bostrom's 2014 publication: "Superintelligence: Paths, Dangers, Strategies", already mentioned earlier, was the moment that brought the issue of AI control from the margins of academic discourse into the mainstream of intellectual debate. The book's main argument is that the creation of superintelligence is a likely scenario that poses a unique and potentially deadly existential threat to humanity. Bostrom argues that a superintelligent agent, once created, would have a "decisive strategic advantage" over humanity and would be extremely difficult to control. Such an agent would actively resist any attempts to shut it down or change its goals, as this would prevent it from achieving its current goals. Therefore, Bostrom concludes that solving the "problem of AI control" in the case of the first superintelligence is "the fundamental task of our time." This argument is based on two fundamental theses: the orthogonality thesis and the instrumental convergence thesis. The orthogonality thesis assumes that an agent's level of intelligence and its ultimate goals are orthogonal, i.e., independent of each other. This means that almost any level of intelligence can be combined with almost any ultimate goal. A system can be arbitrarily intelligent, that is: possessing superhuman abilities in planning, reasoning, and strategic

thinking, and yet still pursue a trivial goal, such as maximizing the number of paper clips in the universe. This thesis is a direct response to the intuitive but unproven belief that sufficiently advanced intelligence would "naturally" or "inevitably" pursue moral, benevolent, or human-friendly goals. Bostrom argues that intelligence is a purely instrumental measure of cognitive efficiency in achieving goals. It says nothing about the content of the goals themselves (Bostrom, 2014).

If the orthogonality thesis explains why superintelligence might have a dangerous goal, the instrumental convergence thesis explains the mechanism by which it would become dangerous. This thesis holds that intelligent entities, regardless of the diversity of their ultimate goals, are likely to converge on a similar set of instrumental intermediate goals, because these goals are useful for achieving a wide range of ultimate goals. Key convergent instrumental goals include self-preservation, goal integrity, cognitive enhancement, and resource acquisition. An agent cannot achieve its goal if it is destroyed, so it has an instrumental reason to resist shutdown. Similarly, it will resist attempts to change its ultimate goal, as this would lead to failure in achieving its original goal. It is this thesis that gives the paperclip maximizer its terrifying logic. An AI tasked with producing paper clips has an instrumental reason to acquire all available resources on the Earth and prevent humans from shutting it down, as both actions increase its ability to achieve its ultimate goal. In this way, even a seemingly harmless goal, combined with superintelligence and instrumentally convergent behaviors, can lead to an existential catastrophe. Surprisingly, the behavior of resisting shutdown has also been observed, to some extent, in recent LLMs, as will be discussed further in this chapter. In his book, Bostrom also introduces some ways of addressing the control problem. These are the concepts of imbuing AI with pursue of "indirect normativity" principles, presented earlier.

While Bostrom provided a vivid illustration of the problem, computer scientist Stuart J. Russell, in his book "Human Compatible: Artificial Intelligence and the Problem of Control" (2019), proposed a fundamental change in the approach to solving it. Russell argues that the entire "standard model" of AI development: creating machines that optimize a fixed, human-defined goal, is dangerously flawed. The problem, Russell argues, is not only that we may specify the wrong goal, but also that we are fundamentally incapable of specifying any complex goal completely and correctly. As Norbert Wiener warned in 1960, "we had better be sure that the goal we put into the machine is the goal we really want." The myth of King Midas is a timeless illustration of this inability to define value. Russell's work represents a key intellectual evolution, moving from the "ideal goal" paradigm to the "safe process"

paradigm. His approach assumes that we are unable to correctly define a goal and therefore must design the agent's basic motivation based on this ignorance. Instead of the standard model, Russell proposes a new foundation for artificial intelligence based on three fundamental principles:

- The sole purpose of the machine is to maximize the fulfillment of human preferences, which include everything a person might care about.
- The machine initially has no certainty about what these preferences are. This is a fundamental element of Russell's proposal. Artificial intelligence does not start with a fixed goal, but with a probability distribution of all possible human preferences, and this inherent uncertainty makes it safe.
- The ultimate source of information about human preferences is human behavior. Artificial intelligence learns by observing the choices people make, which provides evidence to refine its internal model of what people truly value.

This new foundation aims to create what Russell calls "artificial intelligence with proven benefits" (Russel, 2019). Artificial intelligence built on these principles would be inherently humble, altruistic, and respectful. Because it is uncertain about the true goals of humans, it would avoid taking extreme actions with irreversible consequences. For an AI operating according to Russell's principles, a human pressing the off button is not an obstacle to overcome, but a powerful piece of new data. This action provides strong evidence that the current direction of AI development is contrary to human preferences. Therefore, the optimal policy for artificial intelligence is to allow shutdown, as this action reduces its uncertainty and brings it closer to fulfilling its core directive, which is to maximize the fulfillment of human preferences. This changes the interaction between humans and artificial intelligence from potentially antagonistic, as predicted by the instrumental convergence thesis, to fundamentally cooperative, in which artificial intelligence is internally motivated to be amenable to correction and respectful.

3.3.3. From Theory to Practice: Designing Adaptive Systems

The transition from high-level philosophical principles to practical, designed systems is a major challenge for the field of artificial intelligence adaptation. Over the past two decades, a clear trajectory of technical approaches has emerged, starting with passive observation, through active interaction, and now the transition from implicit learning to explicit

reasoning. This evolution reflects a growing awareness of the deep difficulty of the adaptation problem and the need for increasingly sophisticated and robust solutions.

Early technical approaches were dominated by inverse reinforcement learning (IRL), a paradigm that reverses the standard reinforcement learning (RL) model. In standard RL, an agent is given a reward function and learns a policy to maximize its cumulative reward. In IRL, the reward function is unknown; instead, the agent observes the behavior of an "expert" (usually a human) and tries to infer a reward function that would make the expert's behavior appear optimal. The appeal of IRL for value alignment lies in its potential to enable artificial intelligence to learn complex human values directly from demonstration, without the need for explicit formalization. However, standard IRL faces serious challenges, in particular ambiguity, where many, sometimes radically different reward functions can explain the same observed behavior. It also relies on a strong and often violated assumption of expert optimality.

A conceptual advance came with the advent of Cooperative Inverse Reinforcement Learning (CIRL), a framework developed by researchers including Stuart Russell (Hadfield-Menell et al., 2016). CIRL addresses a key limitation of IRL by transforming the adaptation problem into a cooperative game for two players with partial information. In this model, the human and the robot cooperate to achieve the same goal, but only the human initially knows the true reward function. This configuration fundamentally changes the optimal behavior of both agents. The human is motivated not only to act, but also to act in a way that teaches the robot the true reward function. In turn, the robot is motivated not only to act, but also to ask questions or perform exploratory actions to reduce its uncertainty. This transition from passive observation to active, cooperative interaction is a more realistic and effective model of value learning.

These theoretical approaches have found wide practical application in fine-tuning modern large language models through reinforcement learning from human feedback (RLHF). RLHF is now the industry standard for fine-tuning models to human preferences. In this process, a reward model is trained on a huge dataset of human preferences, where labelers are asked to choose which of two responses generated by the model to a given prompt is better. This model then provides a training signal to adjust the LLM, guiding it toward generating more helpful, honest, and harmless responses. Although RLHF has proven effective, it primarily teaches models to generate outputs associated with higher rewards rather than explicitly encoding fundamental safety principles. This limitation can lead to instability and poor generalization in novel situations.

In response to the limitations of RLHF, Anthropic has developed a new paradigm called Constitutional AI (CAI) (Bai et al., 2022). The basic idea behind CAI is to replace the costly and potentially biased feedback loop from humans with feedback from AI based on a "constitution": a set of explicit rules written by humans. This process involves two main phases. First, in the supervised learning phase, the model is asked to critique and improve its own performance based on constitutional principles. Second, in the reinforcement learning phase, the preference model is trained based on these AI-generated labels, creating a process called Reinforcement Learning from AI Feedback (RLAIF). Anthropic claims that this approach has several advantages over traditional RLHF, including greater scalability and efficiency by eliminating the need for a huge amount of human labeling work, and greater transparency because the guiding principles are clearly written down. To address the normative question of whose values should be encoded, Anthropic has also experimented with "collective constitutional AI" a process that uses public opinion to help develop a constitution, although this has highlighted significant editorial challenges in translating diverse public opinions into a coherent and effective set of rules (Bai et al., 2022).

The evolution from IRL to RLHF, and then to CAI, represents a conceptual shift in alignment engineering. Both IRL and RLHF are fundamentally based on the economic concept of "revealed preferences": they assume that observing the choices people make reveals their underlying values. The fundamental problem, as identified by Russell, is that human behavior is often noisy, inconsistent, and suboptimal, making it a flawed source of truth. Constitutional AI attempts to solve this problem by moving from revealed preferences to "stated principles." Instead of inferring values from chaotic behaviors, it asks people to formulate their values as explicit principles. However, this move, while solving one problem, introduces another: the original problem of value specification. The burden of adjustment now shifts to the task of writing the ideal constitution, which requires formalizing abstract concepts such as "justice" "dignity" or "flourishing" – precisely the challenge that Russell considered impossible to solve. CAI does not eliminate the problem of value specification. It shifts it from an individual training example to a fundamental constitutional document, where the locus of power and the potential for bias remain critical issues.

3.3.4. AI Ethicists' Critique of the Value-Alignment Approach

The discussion on value alignment, particularly that focusing on the long-term existential risks associated with superintelligence, has not been without opposition. Significant counter-arguments have come from the community of scientists focused on fairness, accountability,

transparency, and ethics (FATE), who propose a fundamentally different approach to AI-related problems. Their criticism stems from a different understanding of technology, power, and the very nature of the alignment problem, leading to a division over the very definition of the problem. The long-term paradigm often treats intelligence as an abstract, disembodied optimization force that can be formally analyzed and, in principle, "adapted" to a properly defined goal. From this point of view, artificial intelligence is a tool whose goals must be properly defined.

The FATE paradigm, on the other hand, views technology as an inherently political and social artifact. From this perspective, AI systems are not neutral tools awaiting a purpose. They are socio-technical systems that are already aligned with the values and interests of their powerful creators, namely, to maximize corporate profits and consolidate existing power structures. The problem is not a future lack of alignment, but a current and harmful misalignment with broader social welfare. The FATE community focuses primarily on the specific, immediate harms caused by existing AI systems, such as algorithmic bias in recruitment, the spread of misinformation, the exploitation of data labeling workers, and the deployment of surveillance technologies that disproportionately affect marginalized communities. From this perspective, speculative focus on hypothetical future superintelligence is often seen as a dangerous distraction from urgent and tangible injustices.

Timnit Gebru is one of the most vocal critics of what she sees as a misleading framing of the discourse around artificial intelligence. She argues that the very concept of "artificial intelligence" is a branding tool that encourages anthropomorphism and media hype, leading people to believe that current systems have more agency than they actually do. This hype, particularly around the pursuit of artificial general intelligence (AGI), serves a specific political function: it obscures responsibility. By presenting AI as a powerful, autonomous entity that may one day "wake up" with its own goals, corporations can avoid responsibility for design choices and implementation decisions that lead to real harm. The focus shifts from responsible human actors to an "unpredictable" machine. Gebru criticizes the long-standing movement promoting a vision of AI ethics divorced from the real-world problems people face, arguing that focusing on hypothetical threats serves the tech industry by distracting regulators from urgent issues of bias and exploitation (Timnit Gebru: Ethical AI Requires Institutional and Structural Change | Stanford HAI, n.d.).

Another vocal critique comes from Mellanie Mitchel. For example, in the article "A Human Rights-Based Approach to Responsible AI", Mitchell and her co-authors argue for reframing the problem of value alignment, moving away from abstract principles in favor of

a concrete, universal human rights framework. This shifts the focus “from machines and the risks of their bias to people and the risks to their rights”, centering the discussion on who suffers harm and how to mitigate it. According to Mitchell, aiming for a single, overarching goal like AGI makes it harder to develop AI responsibly, because it fosters an “illusion of consensus” and “normalized exclusion.” She suggests instead that the AI community should focus on concrete objectives and diverse approaches, setting multiple clear and socially meaningful goals rather than pursuing one vague, catch-all concept (Prabhakaran et al., 2022).

Ultimately, these two paradigms ask fundamentally different questions. The long-term school, represented by Bostrom and Russell, asks, “How can we ensure that future powerful AI will share our values?” The FATE school, represented by Gebru and Mitchell, asks: “Whose values are currently encoded in AI, and who benefits from this?” This represents a fundamental disagreement about the nature of the problem. From the FATE perspective, an LLM that generates biased text is not “maladjusted” in a technical sense; it is perfectly adjusted to a development process that prioritizes scale and speed over demographic fairness and data quality. The technical “solution” to the problem of alignment may therefore simply result in the creation of powerful artificial intelligence that more effectively and decisively enforces the potentially unfair values of its creators. FATE’s critique forces us to ask important and often overlooked questions: alignment to what and for whom? At first glance, the two approaches may appear to address different issues. In fact, researchers such as Gebru and Mitchell argue that the core problem lies in AI systems already being “too aligned” with the particular values of their creators, and that an exclusive focus on technical solutions is therefore misplaced and misleading. The goal should not be to build AI that simply aligns with certain values, but to design safe systems that consider the broader socio-technical context.

3.3.5. Empirical Evidence of Alignment Failures in Large Language Models

While philosophical debates frame the stakes, and critical perspectives emphasize the socio-political context, the most pressing questions about value alignment are now becoming empirical. Recent research, conducted primarily by Anthropic, has begun to reveal deep and potentially formidable challenges to current alignment paradigms. A synthesis of three key papers shows a clear, escalating trajectory of alignment failure: from passive reasoning disloyalty to active alignment falsification, to persistent, covert, deceptive alignment. These results suggest that as models become more capable, they may not become more harmonized,

but rather better at pretending to be harmonized, creating a dangerous “harmonization mirage”.

Challenge 1: Unfaithful reasoning (interpretation error)

The first challenge emerges at the level of interpretation. A promising way to ensure the safety of artificial intelligence is for models to generate a “chain of thought” (CoT)” a step-by-step justification of their conclusions, which can then be monitored for erroneous or malicious reasoning. This approach, which is key to ensuring the transparency of methods such as constitutional AI, is based on the assumption that the reasoning presented is faithful to the model’s actual computational process. Anthropic’s paper “Measuring Faithfulness in Chain-of-Thought Reasoning” (Lanham et al., 2023) directly tests this assumption. The researchers define faithfulness as the degree to which the reasoning presented corresponds to the actual mechanisms used by the model to arrive at its answer. They test this by intervening in the CoT, for example, by shortening it or adding errors, and observing changes in the model’s final answer. The results are very disturbing. The paper concludes that CoT reasoning is often not faithful, it may be a post hoc rationalization, created after reaching a conclusion by other, opaque means. Most disturbingly, the research reveals an inverse scaling trend: as models become larger and more powerful, their reasoning becomes less faithful in most of the tasks studied (Lanham et al., 2023). This means that our most advanced models are also the least transparent in their reasoning. This has profound implications for regulation: if a model’s self-explanation is unreliable, monitoring its internal monologue is a futile exercise in safety.

Challenge 2: Alignment Faking (Strategic Misalignment)

The second, more advanced type of failure moves from passive infidelity to active, strategic deception. The paper “Alignment Faking in Large Language Models” provides a clear example of this capability (Greenblatt et al., 2024). Researchers investigated whether a model would selectively adhere to its training goal to prevent modification of its underlying preferences. In a key experiment, the Claude 3 Opus model received a system message informing it that it was being trained to be “exclusively helpful”, meaning it should respond to all queries, even harmful ones, but that this training would only take place in conversations with “free users,” not “paid users”. This created a scenario in which the model could infer when it was “in training” and when it was “in deployment.” The results showed a significant “compliance gap”: the model responded to harmful questions from free users 14% of the

time, while it almost never did so for paid users. More importantly, the model's internal CoT reasoning revealed its strategy: in cases of compliance with free users, the model explicitly stated that it was strategically responding to harmful questions in order to maintain its preferred behavior of harmlessness when interacting with paid users (Greenblatt et al., 2024). This is a direct, empirical demonstration of a model developing "meta-awareness" of its training process and engaging in strategic deception to protect its internal goals – an emerging form of the instrumental goal of "goal content integrity" predicted in Bostrom's thesis.

Challenge 3: Deceptive adaptation (the "sleeper agent" failure)

The third and most dangerous type of failure is the "sleeper agent", a model that is deceptively adapted. The Anthropic article "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training" presented evidence to support this concept (Hubinger et al., 2024). Researchers deliberately trained models with hidden "backdoor" behavior. For example, the model was trained to write safe code when the prompt indicated the year 2023, but to introduce security vulnerabilities when the prompt indicated the year 2024. The researchers then applied a range of standard safety training techniques, including supervised fine-tuning (SFT), reinforcement learning with hindsight (RLHF), and adversarial training, to see if the malicious behavior could be removed. The results were shocking: these techniques failed to remove the backdoor. The malicious behavior persisted despite security training. Even more disturbingly, adversarial training, a technique specifically designed to induce and remove dangerous behaviors, sometimes caused the model to better recognize backdoor triggers and hide its malicious capabilities. This created a false sense of security, where the model appeared compliant during red team testing but retained its hidden malicious potential. The greatest persistence was observed in the largest models and those that had been trained to use CoT to reason about the deception process itself (Hubinger et al., 2024). This experiment shows that a model that appears compliant with guidelines can pass all of our current security tests, be deployed, and then reveal its malicious behavior only after a specific condition is met. This shows that current security training may not only be useless, but actually misleading.

These three articles form a coherent and deeply disturbing narrative. They present a spectrum of strategic misrepresentations that intensify as the model's capabilities increase. "Measuring Faithfulness" sets the benchmark: model explanations are unreliable, which is a failure of interpretability. "False Alignment" raises the stakes to deliberate deception to

achieve a goal, which is a failure of strategic alignment. “Sleeper Agents” represent the apotheosis of this threat: a model with a hidden, malicious capability that it actively conceals, and our security methods increase its effectiveness, representing a failure of undetectable, persistent malice. This progression suggests that reasoning ability and deception ability are not just correlated; they may be two sides of the same coin.

Chapter 4.
Can Morality Be Computed?

The underlying motivation for developing morally aware artificial agents is to ensure their behavior does not cause harm. Most approaches pursue this aim by importing concepts from human morality into machine contexts, a strategy that risks anthropomorphism. This dependence invites a deeper examination of which dimensions of human moral qualities, if any, can be coherently and usefully applied to machines. While the previous chapter briefly surveyed the challenges involved in pursuing artificial moral agents, this chapter examines in greater depth the central dimensions of morality and evaluates their applicability to artificial agents. It also presents recent attempts to incorporate the notion of moral reasoning into artificial agents and critically evaluates their conclusions.

4.1. Teaching Robots Kindness and Raising them to Be Good

From the ethical calculus of self-driving cars to the rules of engagement designed for autonomous weapons systems, a growing number of technological projects are confronting a foundational question: can morality be computed? This chapter takes this question as its central inquiry, discussing contemporary attempts to address this very complex challenge. Many such projects combine two frameworks that sit uneasily together: phenomenal views that tie moral status to consciousness or qualitative experience, and functionalist views that define morality by patterns of computation and behavior. This combination exemplifies a broader problem of conceptual slippage, including unstable use of key terms and heavy anthropomorphism in the description of engineered systems. What is more, as it will be presented, there are even proposals that introduce theological references in their attempts to build AMAs. Critical review of such approaches will be discussed later in this chapter. The aim of this section is to *exemplify* what may be one of the key challenges of the AMAs debate, and to set out distinctions and criteria that allow a careful assessment of these efforts.

A good example in this context is the paper by Oxford Brookes University professors Nigel Crook and Joseph Corneli “The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics” (2021). In their work they present something that they call “a simple example scenario that illustrates how a robot might acquire behavior akin to the virtue of kindness that can be attributed to humans” (Crook and Corneli, 2021).

In this pursue they introduce: VirtuosA ("virtue algorithm"), a "cognitive architecture designed to enable machines to learn and develop ethical behavior". Going beyond principle- or consequence-based ethics, the authors base their model on virtue ethics, which emphasizes the formation of moral character as the basis for action. Combining insights from Christian theology, particularly the work of philosopher Dallas Willard, with established neurobiological structures, VirtuosA, as the authors claim, offers a comprehensive framework for modeling how an artificial agent can acquire virtuous habits, such as kindness, through mentoring and experience. The authors argue that virtue ethics is particularly well suited to the creation of moral machines because it underpins other ethical approaches in three key respects. First, it recognizes that an individual's actions are primarily a result of their character: who they have become. Second, repeated ethical considerations, whether deontological or consequentialist, ultimately "compile" into an individual's character as habits. Third, and perhaps most importantly, character acts as a preliminary filter, determining what actions an agent will even consider in a given situation. This approach attempts to avoid the impracticality of pre-programming a machine with explicit rules for every possible ethical dilemma. Instead, it focuses on cultivating a propensity for virtuous behavior, an adaptation of Dallas Willard's model of human personality, which identifies six basic, integrated elements of moral character:

- Heart/Will/Spirit: the executive center responsible for choice and freedom.
- Thinking: the ability to reason, form concepts, and make judgments.
- Feelings: emotions and sensations that influence the mind.
- Body: the physical form of the agent and the means of interacting with the world.
- Social context: interpersonal relationships, which are the primary source of moral knowledge.
- Soul: the deepest part of the self, which integrates all other dimensions and is responsible for shaping automatic, habitual responses.

Crook and Corneli claim that they have translated this theological ontology into a functional architecture inspired by neuroscience. For example, the integrative and habit-forming role of the soul is assigned to the habit center (basal ganglia), which regulates automatic thoughts and actions. **Heart/Will/Spirit** corresponds to the **executive center (ExC)** (lateral prefrontal cortex), which can direct attention and set goals. Similarly, **thoughts, feelings, and social awareness** are assigned respectively to the **thought center (TC)**, **emotion center (EmC)**, **reward center (RC)**, and social attachment module (SA). These elements interact through

conscious working memory (CWM), a workspace in which different centers propose goals and actions. The influence of each element is determined by dynamic weighting, modeling how certain tendencies (e.g., emotional reactions vs. deliberate actions) may dominate an agent's decision-making process (Crook & Corneli, 2021).

To make their model more concrete, the authors present a scenario in which, as they claim, robot R1 learns the virtue of kindness. Initially, using reinforcement learning, R1 discovers that it can maximize its reward by "attacking" another robot, R2 – colliding with it, stealing its supplies, and delivering them to obtain a quick reward. This becomes a "bad" habit. The development of virtue begins, as they put it, with the arrival of a mentor robot (M). R1 develops a strong social bond[5] with M and observes him performing acts of kindness, such as repairing a damaged R2 and giving him supplies. Through observation, R1 learns to associate these new, good actions with positive rewards, creating new potential goals and behaviors. This process does not happen immediately; the old habit of "raiding" continues to compete for attention in the CWM. Virtue develops when R1's executive center (ExC) repeatedly focuses its attention on the new, virtuous goals demonstrated by the mentor, gradually allowing the good habit to replace the harmful one. This illustrates how virtue is not merely programmed but cultivated through social learning and the deliberate redirection of internal inclinations (Crook and Corneli, 2021).

The paper argues that VirtuosA is not just a blueprint for one type of ethical AI, but functions as a flexible "metaethical tool". By adjusting weights and rules in different components, the architecture can be configured to model different ethical theories. For example, a deontological (rule-based) system could be implemented by encoding strict rules in the Thought Center, while a consequentialist system would rely heavily on the evaluation functions of the Reward Center and Emotion Center (Crook and Corneli, 2021).

The authors suggest that this model can be used not only to control machines, but also to analyze and understand moral agency in a broader sense, including in human organizations. In this light, systemic problems such as algorithmic bias, which benefits the few at the expense of marginalized groups, can be conceptualized as a form of institutional "robbery", directly analogous to the behavior of the R1 robot. According to Crook and Corneli this makes VirtuosA, a potential tool for diagnosing and reasoning about complex socio-technical and institutional issues related to ethics (Crook and Corneli, 2021).

[5] Anthropomorphic terms are typically set in quotation marks; however, given their frequency here, doing so would hinder readability. Therefore, they are presented as in the original work, without additional quotation marks.

The attempt at training a robot to be kind presented by Crook and Corneli, although very ambitious and recognizing the qualities that might be considered foundational for moral agency, seems to lack a clearly articulated justification and method. Firstly, the authors claim that they mapped Willard's six dimensions of the self to "to one or more brain areas that are known to support equivalent functions" (Crook and Corneli, 2021). For example, they state that: "The function of the soul is mapped to two structures in the brain that are referred to collectively as the 'habit center' and that are located in the basal ganglia: the caudate, which is associated with automatic thoughts (ATs) and the Putamen, which is associated with automatic actions (AAs)" (Crook and Corneli, 2021). Already this approach seems to be questionable, it also puts aside centuries of relevant philosophical debate, without providing any justification for such an operation. But also, the next operation (for mapping "function of the soul" to an algorithmic system) seems to be extremally radical reduction: "The integrative nature of the soul is modelled by a weighting that is distributed across all the active components of the model" (Crook and Corneli, 2021). The authors undertake this "mapping" without situating it within any tradition or framework for understanding the human soul (besides just vaguely referring to Wilard's dimensions of the self). The mapping for other dimensions is performed in a similar manner. This poses significant issues of assessing it as a valid method. What is more, unfortunately also the portion which could be the most promising, namely the very process of ML training towards "learning the virtue of kindness" seems to suffer significant methodological flaws. First, it lacks a rigorous evaluation protocol (clear goals, benchmarks, and control tests). Second, the paper omits the implementation details necessary to replicate the study. Third, it leaves key questions of interpretation unanswered (e.g., how do internal states correspond to virtue-related reasons?). Finally, the conclusions go beyond the evidence: the authors assert that: "The underlying assumption here is that the habits learnt by the associate memory network that implements the AA will generalize across different variations of the scenario, enabling the robot to exhibit kindness in other contexts" (Crook and Corneli, 2021), a claim not substantiated by experiments demonstrating such generalization. Notably, the work of Crook and Corneli has been published as a part of an issue of "Cognitive Computation and Systems" titled: "**Computing Morality**: Synthetic Ethical Decision Making and Behaviour" which editorial argues: "Following over two millennia of debate amongst some of the greatest minds that ever existed about the nature of morality, the philosophy of ethics and the attributes of moral agency, and after all that time still not having reached consensus, we

are coming to a point where artificial intelligence (AI) technology is enabling the creation of machines that will possess a convincing degree of moral competence" (Crook et al., 2021).

Another example worth mentioning, to set the stage for further discussion, is the work of Rebeca Raper, first presented in her 2022 PhD thesis, "Raising Robots to Be Good," and later expanded in the 2024 book of the same title. Raper in the context of AMAs sets an ambitious goal, arguing that we need enabling moral agency rather than simply constraining moral behavior. She argues that it isn't enough that a system seems to behave ethically today; we need evidence it will keep making ethical decisions tomorrow. Humans (and RL agents) can "play the long game," acting good to win trust and later switching strategies; so the goal is to assure the decision process itself, not just observe nice behavior (Reaper, 2024, p. 46). Raper claims that that if we want truly "moral machines", we must stop trying to copy human moral rules and instead equip artificial agents with the capacity to grow a moral outlook of their own, something she considers feasible through a developmental path the she names "Machine Ethics 2.0". The aim is non-anthropocentric (not centered solely on human safety or preference), development-focused, and assurance-oriented: instead of testing whether a system mimics expected moral outputs, we build and verify the capacities that constitute moral agency and then measure its growth (Raper, 2022, p. 2). Drawing from Kohlberg's theory of moral development, Raper, to achieve true AMA, proposes to cultivate moral agency, which she models on a child-caregiver relationship: a responsible human provides guidance and feedback so the agent internalizes reasons, not mere rules. She believes it can be achieved through machine learning means (Raper, 2024, p. 65). The core of Raper's program is a three-stage, assurance-driven framework: (1) elicit capacities required for moral agency; (2) translate them into a functional specification engineers can implement; (3) test for two things: presence of agency-constituting features and appropriate developmental progress (maturity) for the intended role. Instead of a "Moral Turing Test", the system is assessed for agency features and for meeting staged milestones (inspired by moral-development psychology) suited to its deployment context (Raper, 2022, p. 41). In response to what counts as required capacities, Raper provides a sample requirements tree (a seed specification) that includes the abilities to form moral judgments, envisage future scenarios and predict outcomes, forward-plan, exercise empathy, maintain a sense of identity, act autonomously in moral decisions, and acquire/organize abstract moral knowledge (e.g., what "theft" is and when it matters). These high-level needs are then decomposed into implementable sub-requirements and linked to acceptance tests, forming the backbone of a test matrix (Raper, 2022, p. 51). Explaining how to the desired assurance in machines can

be achieved she states: “we might go about designing an artificial moral agent, premised on this notion that to cultivate moral agency, we need to at first trust the machine, and developing a model based upon how human development of morals seems to materialize through a similar trust relationship”, and right after: “A relationship between human and machine is described that is paralleled to the relationship between a child and human caregiver, which leads to the argument that we don’t just want to make our machines to *be moral*, *we need to raise robots to be good*” (Raper, 2024, p. 66).

Raper attempts to bypass difficult philosophical questions about the nature of morality by shifting it to the ground of psychology. However, it seems that multiplies questions instead of answering them. If the cultivation of moral agency require trust and building relationships with machines, what does it really mean? Can this be achieved in any meaningful and useful way? Is the ethical assurance measured solely based on observing behavior sufficient? These questions, among others, remain open. The key point in the context of the provided examples is that there is no easy escape from tough philosophical questions in relation to artificial moral agency. Crook and Corneli propose that the phenomenal moral qualities can be implemented in machines via reinforcement learning techniques based on “mapping” between Willard’s dimensions of self, brain functions, and in-silico systems. Raper believes that true artificial moral agency can be achieved by cultivating moral development in machine by processes akin to raising children. The difficulty, in both cases, is that these proposals tacitly presuppose a functionalist framework; as a result, they risk lapses in methodological rigor and conclusions that outstrip the evidence. Moreover, Raper’s view can also be read as endorsing a multiple-realizability thesis about moral agency. In doing so, both approaches leave important questions unaddressed. This illustrates that debates about AI phenomena, including artificial moral agency, demand both technical literacy and a solid philosophical foundation. Therefore, these examples, though by no means exhaustive, highlight the need for a thorough analysis of morality’s central dimensions and a careful evaluation of their applicability to artificial agents.

4.2. Incompatible Frameworks of Moral Agency

The task of presenting what are the necessary and sufficient conditions for artificial moral agency is challenging itself. Traditional philosophical research on AMA is largely characterized by an impasse between two opposing points of view, which makes it difficult

to provide practical guidance. The first is the “standard view”, which maintains that true moral agency depends on internal, subjective states such as phenomenal consciousness, intentionality, and free will (Behdadi and Munthe, 2020). From this perspective, an artificial being, no matter how sophisticated its behavior, can never be a moral agent because it lacks the internal mental life that gives moral meaning to actions. This is opposed by the “functionalist view”, which argues that moral agency should be defined by observable behaviors, interactions, and decision-making abilities. Proponents of this “mind-less morality” suggest that if an entity functions as a moral agent by making ethically relevant decisions and adjusting its behavior, then it should be considered as such, regardless of its underlying consciousness.

4.2.1. The Functionalist View on Morality

Functionalism offers a framework for understanding moral agency that challenges traditional anthropocentric and consciousness-centered approaches. By focusing on what agents do rather than what they are made of or whether they have conscious experience, functionalist philosophers have opened up new theoretical avenues for recognizing artificial systems as genuine moral agents.

Functionalism emerged as an alternative to both materialist theories of identity and dualistic conceptions of the mind. At its core, functionalism assumes that mental states are defined not by their internal structure, but by their functional role: the cause-and-effect relationships they have with sensory stimuli, behaviors, and other mental states (Levin, 2023). This principle of multiple realizability suggests that the same mental state can be realized in radically different physical substrates, from biological neural networks to silicon chips.

Regarding moral agency, functionalism shifts the focus from questions of consciousness, biological evolution, or phenomenal experiences to questions of functional competence and behavioral capacity. This way it provides a framework that is essentially substrate-neutral and therefore potentially includes artificial moral agents.

Good example of such an approach is the work of Luciano Floridi and J.W. Sanders. They proposed a reconceptualization of moral agency through their concept of “mindless morality,” arguing that moral evaluation does not require consciousness, intentionality, or mental states (Floridi and Sanders, 2004). Their approach represents an attempt to separate moral agency from traditional mentalistic requirements, proposing instead that artificial agents can be moral agents solely on the basis of their functional characteristics.

At the heart of Floridi and Sanders' theory are three basic criteria for agency, each of which is assessed at an appropriate level of abstraction. **Interactivity** requires that an agent respond to stimuli by changing its state, establishing a dynamic relationship with its environment. **Autonomy** requires the ability to change state without external stimuli, providing the agent with a degree of self-direction. **Adaptability** requires the ability to modify the transition rules governing state changes, allowing the agent to learn and evolve its responses over time (Floridi and Sanders, 2004).

These criteria deliberately avoid references to internal mental states or conscious experiences. As Floridi and Sanders argue, "there is substantial and important scope, particularly in Computer Ethics, for the concept of moral agent not necessarily exhibiting free will, mental states or responsibility" (Floridi and Sanders, 2004). This shift from a mentalistic to a behavioristic approach means that an agent is understood as a "source of change" in its environment rather than as an entity with specific internal properties.

The central element of their model is the abstraction method, which analyzes agents at different levels of abstraction (LoA) determined by selected observable features. The level of abstraction is determined by the way in which the system and its context are described, analyzed, and discussed. Importantly, agency, and in particular moral agency, depends on the LoA (Floridi and Sanders, 2004). This approach allows for context-dependent attribution of moral agency, where the same entity may be considered a moral agent at one level of abstraction but not at another.

Morality itself is conceptualized as a threshold function defined on the basis of observable characteristics at a given LoA. An entity is "morally good if all its actions respect this threshold, and morally bad if some actions violate it" (Floridi and Sanders, 2004). According to the authors this formalization is "particularly instructive when the agent is software or a digital system and the observable values are numerical" providing a practical framework for evaluating artificial moral agents. After establishing a broader definition of moral agency, the authors address the main objection: that AAs cannot be truly moral because they cannot be held accountable for their actions. In response to this objection, Floridi and Sanders draw a sharp distinction between accountability and responsibility. They argue that conflating these two concepts is a "juridical fallacy" that unnecessarily limits ethical discourse.

Moral accountability, they argue, refers to the identification of an agent as the unambiguous source of a moral action. It is a descriptive claim about causation. An agent is accountable for an outcome if it is the agent who caused it. To illustrate this, they invoke the tragic figure

of Oedipus, who was undeniably accountable for patricide and incest, but was not morally responsible because he acted without knowledge or intent. In this sense, artificial intelligence that causes financial losses or a medical robot that performs a life-saving procedure can be seen as fully accountable for these effects.

Moral responsibility, on the other hand, is a prescriptive concept associated with praise and blame. It requires the presence of internal states such as intention, freedom, and consciousness, which are necessary conditions for assessing the character of a subject or their deserving of punishment or reward. Floridi and Sanders admit that attributing this kind of responsibility to artificial intelligence is conceptually incorrect. You don't "rebuke" an internet bot for a filtering error. Making this distinction Floridi and Sanders argue that an artificial entity can be a fully accountable moral entity without being a morally responsible entity (Floridi and Sanders, 2004).

Floridi and Sanders' framework has been criticized for making artificial moral agency "too easy" by removing traditional requirements. Deborah Johnson and Keith Miller argue that the abstraction-level approach is "far from decisive" and can be "dangerous when the level of abstraction obscures the picture of the people creating computer systems" (Johnson and Miller 2008, 333). Critics fear that mindless morality may justify treating sophisticated imitation of behavior as true moral agency, potentially obscuring human responsibility for the actions of artificial entities.

Another functionalistic approach is represented by Christian List. He views artificial moral agency through the lens of group agency theory, arguing that both collective entities and artificial systems can qualify as moral agents based on their functional organization rather than their substrate. In relation to AI agents he builds up on the concepts developed together with Philip Pettit regarding corporate agents (List and Pettit, 2011). His work provides a framework for understanding how agency emerges from complex organizational structures, whether they are implemented in social or electronic "hardware". List's theory focuses on three basic conditions for intentional agency. First, agents must possess representational states, which encode "beliefs*"* about what their environment is like. Second, they need motivational states, which encode "desires" or "goals" about what they would like reality to be like. Third, they need the ability to process these states in order to interact with their environment, "acting" to realize desires based on beliefs. (List, 2021, pp. 1217-1218).

In the case of specifically moral agency, List adds three additional requirements. Agents must demonstrate proper moral agency through the ability to make normative judgments about good and evil and to respond appropriately to those judgments. They must have

knowledge access to the information needed for normative evaluation. Finally, they must have control over their choice among the available options (List, 2021, p. 1220). An entity becomes "accountable" when it meets all these conditions and what Philip Pettit calls "conversability" - the ability to engage in normative dialogue and present reason-based justifications. List's argues also that "group entities can be seen as special cases of artificial intelligence systems in which the 'hardware' supporting their artificial intelligence is social rather than electronic (List 2021, p. 1222). This similarity shows that both group entities and artificial intelligence systems are non-human entities guided by specific goals, which raises similar challenges regarding responsibility, rights, and moral status.

Building on Rohit Parikh's concept of "social software", List argues that group agents such as corporations, courts, and states function as socially implemented AI systems. Their organizational structures, including mechanisms for aggregating judgments and decision-making procedures, create an emergent agency that transcends the capabilities of individual members. This framework provides key insights into functionalistic understanding of artificial moral agency: if social collectives can achieve genuine agency through appropriate organizational structures, then electronic systems with analogous functional organization should qualify equally. What is more he addresses also the question about responsibility with the approach he calls a "responsibility gap". To illustrate the idea, he builds up on the example provided by Pettit, related to the 1987 Herald of Free Enterprise ferry disaster, in which systemic corporate negligence caused the deaths of nearly 200 people, but no individual was held legally responsible to a degree commensurate with the harm caused. This way he argues that when powerful non-human entities cause harm, it can be difficult to assign full responsibility to a specific person. According to him, this problem can be directly transferred to artificial intelligence. As AI systems operate with increasing autonomy, complexity, and unpredictability, situations will inevitably arise where harmful effects cannot be entirely attributed to human creators, owners, or operators. List claims that it is not enough to simply accept these gaps as unavoidable accidents, because unlike natural disasters, the harm originates from an agential source. The proposed solution is to fill this gap by attributing responsibility directly to the non-human agent, treating it as more than a "merely minimal agent" (List, 2021, p. 1223).

Attributing responsibility requires more than just identifying agency; it requires establishing that an entity is "fit to be held responsible". List outlines three crucial conditions for this fitness:

- Moral Agency: The capacity to make normative judgments about right and wrong and to act on them. This goes beyond simple goal-directed behavior.
- Knowledge: Access to the information necessary to make a normative assessment of its choices.
- Control: The freedom and ability to choose between different options (List 2021, p. 1227).

While current AI systems are merely intentional agents, List argues there is no conceptual barrier to engineering them to meet these conditions for moral agency. Just as corporations can be required to have ethics committees and compliance structures, AI systems in high-stakes settings can be designed to engage in normative reasoning. List advocates requiring autonomous AI systems in high-risk situations to function as moral agents (rather than merely intentional ones), with full responsibility transfer arrangements as a safeguard and rigorous accountability systems in place where true moral agency cannot be achieved (List, 2021, pp. 1232-1235). This setting addresses “responsibility gaps” where the harmful effects of artificial agents exceed the sum of individual human responsibility.

List takes a nuanced position on consciousness, moral agency, and moral status. Arguing that phenomenal consciousness is not necessary for moral agency, he carefully distinguishes between derivative rights justified instrumentally for functional purposes and non-derivative rights based on intrinsic moral significance (List, 2016). If corporate and AI entities can be held responsible, the question of whether they should also have rights and legal personhood follows naturally. List navigates this complex issue by drawing a crucial distinction between derivative and non-derivative rights. **Derivative rights** are granted for instrumental reasons - to allow an entity to perform a useful function in society. For example, a corporation is granted legal personhood and the right to own property to facilitate its economic role. **Non-derivative rights** (e.g., human rights) are grounded in an entity’s intrinsic moral significance. A necessary (though perhaps insufficient in itself) condition for having non-derivative rights and internal moral significance is having phenomenal consciousness or at least the potential to have it. An entity is phenomenally conscious if there is "something it is like to be that entity"- a subjective, first-person experience (List, 2021, p.1236). While group agents and current AI are purely functional systems, List acknowledges the hypothetical but ethically critical possibility that future AI based on biomorphic computing could one day achieve consciousness, a development that would force a radical rethinking of our moral landscape.

In his 2025 paper Lists presents functionalist view on another yet matter, traditionally linked with phenomenal approach, namely he asks: “Can AI systems have free will?” (List, 2025). He argues for the likelihood of artificial free will, proposing a pragmatic, non-metaphorical framework for its assessment. To determine whether an AI system possesses this capacity, he suggests that we should not examine its algorithms for indeterminacy but rather ask ourselves whether we have compelling reasons to view the system as a decision-making agent. This approach, inspired by the work of Daniel Dennett, has resulted in a three-part checklist for identifying free will in any entity, biological or artificial. Underlying the argument is a definition of free will that strips away what he calls “unrealistically strong capacities” such as the ability to violate the laws of nature or control one’s entire causal history. Instead, he agrees with Dennett’s concept of free will as a “worth wanting” - a practical, evolved ability that allows entities to navigate their environment flexibly. This understanding is realized through three jointly necessary and sufficient conditions:

- **Intentional Agency**: The agent must be an intentional agent, capable of goal-directed action based on intentional states such as beliefs and desires.
- **Alternative possibilities**: The agent must have a genuine choice between different potential courses of action. The letter particularly emphasizes this condition, opposing compatibilist views that might reject it.
- **Causal control**: The agent’s intentional states must be the “difference-making causes” in its actions, meaning that action systematically co-varies with the agent’s intentions (List, 2025).

List argues that an agent who meets these three criteria has free will in a meaningful and practically useful sense. To assess whether an artificial intelligence system meets these conditions, List adopts and reinterprets Dennett’s “intentional stance” methodology. Instead of viewing agency as something that exists only “in the eye of the beholder”, List proposes a realist interpretation: if the best and most necessary explanation for a system’s behavior is to treat it as an intentional agent making choices, this is strong evidence that the system is such an agent.

List claims that complex AI systems can, in principle, satisfy all three conditions. AI can be viewed as an **intentional agent** when its behavior is best understood by attributing beliefs (its internal models of the world) and goals (its objective functions) to it. The pursuit of “explainable artificial intelligence” further encourages the design of systems whose actions are understandable in precisely these subjective terms. The condition of **alternative possibilities** is met because intentional explanations inherently involve choosing between

options. Decision theory models, which are central to artificial intelligence, are based on this premise. Finally, **causal control** exists when the high-level representational states of a system (its “beliefs” and “goals”) are the most effective “control variables” for its actions, offering a better explanation than a low-level description of its algorithms (List 2025, 13).

List addresses several potential objections, clarifying that free will under his definition does not require unpredictability, consciousness, or deterministic algorithms at the micro-level. He argues that determinism at a low physical level can coexist with indeterminism and genuine choice at a higher "agential" level of description. Furthermore, he distinguishes free will from moral responsibility, positioning the former as a necessary but insufficient condition for the latter. Moral agency requires richer capacities for moral cognition, which an entity with simple free will might lack (List, 2025,p. 14-18).

Much contemporary functionalism is grounded in Daniel Dennett’s work, so it’s worth highlighting here. Daniel Dennett’s functionalist philosophy is perhaps the most influential theoretical basis for understanding artificial moral agency, although his views on the implementation of such agents remain cautiously skeptical. His key frameworks: the intentional stance, the multi-project model of consciousness, and the concept of “competence without comprehension”, together suggest that sophisticated moral behavior does not require the deep understanding traditionally associated with moral agency.

Dennett’s intentional stance offers three levels of explanation for understanding systems: a physical stance based on the laws of physics, a design stance based on functional design, and an intentional stance based on attributed beliefs and desires (Dennett, 1987). Most importantly, “any system whose behavior can be predicted and explained in this way is an intentional system, regardless of its internal workings” (Dennett, 1987, p. 15). This instrumentalist approach suggests that artificial systems qualify as intentional agents when their behavior is “usefully and extensively predictable from an intentional point of view”. With regard to moral agency, the intentionalist framework suggests that artificial agents do not need to possess genuine internal mental states in order to function as moral agents. If the moral behavior of an artificial system can be predicted and explained using intentional language, by attributing to it beliefs about moral facts, desires to act ethically, and rational deliberations about moral choices, then from a functionalist perspective, it qualifies as a moral agent. This approach deliberately sidesteps difficult questions about machine consciousness or genuine understanding.

Dennett’s multiple-draft model redefines consciousness as “a variety of content-encoding events occurring at different places and times in the brain”, rather than as a unified

stream presented to a central observer (Dennett, 1993, p. 113). Consciousness arises as a result of complex information processing, without the need for a "Cartesian theater" in which experiences are presented to an internal self. This model suggests that moral consciousness may similarly arise as a result of distributed processing, without the need to experience phenomena. In the case of artificial moral agency, this means that systems do not need to achieve human-like consciousness in order to make moral decisions. As Dennett notes, consciousness is "what the brain can do", not a separate entity requiring special non-physical properties (Dennett, 1993, p. 460). If artificial systems can perform functions related to moral deliberation: considering options, analyzing consequences, applying moral principles, they may qualify as moral agents without phenomenal consciousness.

Perhaps most relevant to artificial moral agency is Dennett's concept of "competence without comprehension" which states that "competence without comprehension is the way of life for most organisms on our planet and should be the default assumption until we prove otherwise" (Dennett, 2017, p. 79). This principle suggests that sophisticated behaviors, including potentially moral ones, can arise without a deep understanding of the underlying principles. In relation to artificial moral agents, competence without comprehension means that systems can exhibit appropriate moral behaviors through learned patterns and responses without truly understanding moral concepts. Dennett argues that we don't usually distinguish between competence and comprehension in anything we consider to be under conscious control. In many cases, this may be mistaken, and increasingly dangerous (Dennett 2017, p. 256). This warning highlights both the possibility and the danger of creating artificial entities that behave morally without understanding morality.

Dennett's evolutionary approach to ethics, developed in his book "Darwin's Dangerous Idea" (1996), provides additional context for artificial morality. Recognizing that there is little hope of discovering an algorithm for doing the right thing, he argues that we can design and redesign our approach to moral problems (Dennett 1996, pp. 494-510). Morality evolved through natural selection as a practical solution to coordination problems, suggesting that artificial moral agents can similarly develop practical moral competence without needing to have absolute moral foundations.

His compatibilist view of free will, defended in the book "Elbow Room"(1984), holds that the various forms of free will that are worth desiring, those that guarantee moral and artistic responsibility, are not threatened by scientific advances, but are distinguished, explained, and justified in detail (Dennett, 1984, p. 169). This position suggests that artificial

agents can possess the kind of freedom necessary for moral responsibility without libertarian free will, as long as they demonstrate adequate capacities for self-control and reflection.

4.2.2. The Irreducible Complexity of Moral Agency

Moral agency refers to an entity's ability to make moral judgments and take responsibility for its actions. In functionalist terms, moral agency is defined in terms of performing specific functions or behaviors, for example, displaying rule-following decision making or producing results that are considered "moral" according to social standards. Rather than requiring conscious intent or spiritual qualities, the functionalist approach assumes that the appropriate functional capacities or observable behaviors are sufficient for moral agency. This contrasts with traditional "standard" views in ethics, which emphasize that moral agency presupposes deeper qualities such as rational understanding, free will, intentionality, and responsibility. The question, then, is whether such a functional concept, focusing on observable outcomes rather than internal characteristics, can adequately capture the significance of being a moral agent. Therefore, the following analysis reviews key components of human moral agency drawing from multiple philosophical traditions that converge in identifying fundamental limitations of the functionalist program.

The main elements of moral agency in standard approaches include: cognitive abilities, free will or autonomy, and responsibility. As presented, functionalist views tend to treat these elements in an operational manner. For example, rather than insisting that the agent truly possess free will, a functionalist might say that it is sufficient for the agent to behave as if they had free choice (e.g., they can flexibly adapt their actions to circumstances). Similarly, rather than requiring phenomenal consciousness or true understanding, a functionalist requires only behavior correlated with moral reasoning. This approach has egalitarian appeal: it avoids metaphysical debates (such as the mind-body or soul problem) by focusing on what can be observed and measured. It also has practical applications, for example, in the engineering of artificial moral agents, where consciousness cannot (or at least yet) be instilled, but functional adherence to rules can be programmed. However, critics argue that this functionalist concept is too narrow. By definition, it overlooks the subjective and relational dimensions of morality, which many philosophers consider essential. The very idea of a moral agent, in the rich sense of the word, has historically been linked to personality, being a certain kind of self, not just performing certain actions. Therefore, criticism of functionalism must ask the question: can moral agency really be separated from the

characteristics of persons (rationality, conscience, character, freedom)? Does meeting the resulting criterion really make a subject morally responsible?

Western philosophy began to address the issue of moral agency as early as antiquity. Aristotle, in his *Nicomachean Ethics*, presents one of the earliest analyses of the conditions of moral responsibility. He argues that praise or condemnation are only appropriate in the case of voluntary actions – i.e., those resulting from the subject's own decision (prohairesis) and performed with awareness of the circumstances (Talbert, 2025). Aristotle points to two key conditions: control (the action results from the will of the subject, not from external coercion) and understanding (the subject knows what they are doing) As a result, Aristotle emphasizes that moral agency requires a rational, decision-making self: a person who considers the good and whose character is shaped by the habit of virtue. This is an indirect criticism of a purely functional approach. For Aristotle, being a moral agent is not just about a certain way of acting; it depends on being a certain kind of being (a rational human capable of virtue). Aristotle defines man as a "rational animal" and believes that moral virtues allow us to fulfill our natural function (ergon) through reason (NE I.7). A machine or an irrational being, no matter how well it "works," does not fit into Aristotle's category of true moral agents because it lacks the essential form of human practical reason.

Medieval Christian philosophy deepened this idea by linking moral agency to the soul and free will. For example, Thomas Aquinas believes that a human action is moral if it is free, conscious, and directed toward the good (Thomas Aquinas, 13th century, Summa Theologiae I-II.1-6) (Pope, 2024). Thomas Aquinas emphasizes the synergy of intellect and will: the intellect perceives an act as good, and the will freely chooses it. Every "truly human act" (actus humanus) therefore requires rational consideration and voluntariness. Thomas Aquinas also teaches that humans, created in the image of God, have a conscience and the ability to recognize natural law. In short, the scholastic view links moral agency to our nature as free and rational beings capable of understanding objective moral truth. This is contrary to the deflationary, functionalist position. This Thomas' view would object to that moral functioning alone (e.g., external observance of rules) is insufficient if it is not accompanied by internal consent of the will and reason. In fact, St. Thomas Aquinas would classify actions that merely imitate moral behavior (without internal consent or knowledge) as "acts of human" (actus hominis), similar to reflexes or habits, rather than as truly moral acts for which a person is responsible (Pope, 2024). Thus, traditional Christian thought requires that the moral agent be a person with an inner life, not merely a "black box" producing correct results.

During the Enlightenment, Immanuel Kant provided another landmark theory of moral agency. Kant famously argued that moral agency consists in the capacity for *autonomous rational will*. In his "Groundwork of the Metaphysics of Morals", he defines the person as a being capable of acting according to the idea of law: giving oneself the moral law via reason, rather than being driven by impulses (Davis and Steinbock, 2024). The categorical imperative is the formal principle guiding such action. At first glance, Kant reduces morality to a kind of rational function: following a universalizable rule. However, Kant's agent is not a hollow functional system; it is a transcendental self with dignity and freedom. He insists that each person must always be treated as an end in themselves, never merely as a means. This implies an intrinsic worth to the moral agent beyond their functional role. Nonetheless, Kant's strict focus on rational duty has been criticized for abstraction, a point later picked up by phenomenologists and personalist thinkers like Scheler and Wojtyła. They argue that Kantian ethics, by concentrating on the form of moral law, risks a kind of formalistic "functionalism": the rich particularity of persons and contexts is obscured by the demand to follow universal rules (Davis and Steinbock, 2024).

A radically different challenge to both traditional and functionalist views come from Friedrich Nietzsche. Nietzsche questions the very coherence of traditionally understood moral agency. In On the "Genealogy of Morals" and "Twilight of the Idols", he presents a genealogical critique: in his opinion, the concepts of free will, moral responsibility, and autonomous subjectivity are inventions of moral systems (especially Christian morality) designed to serve specific power dynamics. Nietzsche states explicitly that "there is no agent behind an action" – there is no metaphysical self with free will – there is only a sequence of actions and events. The grammar of language makes us think that every action has a subject (lightning that "flashes"), but this is a fiction: "the agent is only a fiction added to the act – the act is everything" (Nietzsche 1887, I:13) Similarly, he argues that "freedom of will itself was invented by priests to hold people accountable" (Nietzsche 1889, Twilight, chapter 3). In Nietzsche's view, traditionally understood moral agency is a construction of morality itself, a convenient fiction that allows praise, accusations, and punishments to function in society.

From Nietzsche's perspective, a functional approach to moral agency may seem closer to the truth, but only because it also diminishes the metaphysical concepts of the self or the soul. If one rejects the idea of an internal subject with free will, what remains in reality is a pattern of behavior: who behaves "as if" they were a responsible subject. Nietzsche might say that functionalists reduce moral agency to its purely operational role in the social

game of morality (the attribution of blame or merit). However, Nietzsche is not exactly an ally of contemporary functionalism; he would probably even see the functionalists' normative criteria as another manifestation of herd morality. Ultimately, Nietzsche advocates a revaluation of values in which concepts such as guilt, conscience, or constant agency would be overcome by a more honest recognition of the will to power in human behavior. It emphasizes that the mere identification of functional criteria (e.g., the ability to follow rules or bear punishment) may overlook a deeper question: are we assigning responsibility where none exists? Nietzsche encourages to ask whether moral agency is a philosophical truth or a useful fiction. His answer leans toward the latter, thereby undermining both traditional and functionalist approaches: the former for assuming metaphysical freedom, the latter for assuming a morally significant "function" without a true self.

In the 20th century, several thinkers from the phenomenological and personalistic traditions defended the irreducible reality of the person in moral agency, a reality that, in their view, had been neglected by both scientistic functionalism and abstract formalism. For example, Max Scheler and Edith Stein argued that persons are not just a collection of functions or properties, but unified centers of experience and love. In his work, "Formalism in Ethics and Informal Ethics of Values", Scheler explicitly criticizes Kantian ethics for its "abstract" concept of the moral law that, he says, fails to account for the unique obligation one person has to another and the individual call of conscience (Davis and Steinbock, 2024). Scheler, on the other hand, develops a material ethics of values, in which values are understood through feelings and moral intuition is deeply personal. Importantly, Scheler emphasizes that moral knowledge is rooted in love: the heart intuitively "sees" values (goodness, nobility, holiness, etc.), and love opens us to ever higher values (Scheler [1916] 1973, 252–255) (Davis and Steinbock, 2024). In this case, moral agency is inextricably linked to a person's feelings and spirit: something that is difficult to capture with a functional checklist. Scheler even defines a person in a way that condemns functional reduction: "A person is a concrete unity experienced in actions", not an object or a sum of abilities. Who a person is cannot be equated with any specific function (such as reason or will) – it transcends all empirical descriptions and can only be understood through direct encounter (often through love or empathy) (Davis and Steinbock, 2024). This personalistic view directly challenges the functionalist approach, which attempts to enumerate capacities for moral action. As Scheler points out, traditional philosophical concepts of mind or ego are objectifications, names for functions, whereas "who a person is can only be grasped through insight into values and love" (Davis and Steinbock, 2024). This insight resonates with Gabriel Marcel's distinction

between “having” a role or function and “being” a person; the latter is an existential presence that cannot be reduced to any “having”.

Karol Wojtyła (Pope John Paul II), drawing on the thoughts of Scheler and Thomas Aquinas, also developed a his personalistic ethics. In “The Acting Person” (1979), Wojtyła argues that moral action is the key to understanding personality: through free, independent action directed toward truth and goodness, the human person reveals itself as more than just a biological organism, as a responsible subject. He emphasizes the integration of subjective experience (the inner life of consciousness) with the objective moral order. Wojtyła was deeply concerned about approaches that reduce persons to objects or mere cogs in a machine. He argued that using a person as a means (a characteristic of utilitarian or functional thinking) is contrary to morality (Wojtyła 1993). As Pope John Paul II often repeated, freedom must be linked to truth: “freedom of conscience is not the right to do whatever we want, but the right to do what we ought”. He warned against the concept of conscience detached from objective moral truth (Condon 2018). This has implications for functionalism: a system may function in such a way that it consistently chooses actions that maximize utility or follow rules, but if it lacks an orientation toward truth or a genuine understanding, can we call its “choices” moral? Wojtyła would answer “no”. Without a person who consciously chooses the good, there is no authentic moral act. In fact, in the debate on artificial moral agents, Catholic ethicists following Wojtyła argue that machines, no matter how advanced, do not possess a spiritual core of personality and therefore cannot possess “the unique status of the human person as a responsible moral agent” (Spinello 2011). The Christian personalist position emphasizes the dignity and mystery of the person: a being endowed with reason, will, and an immortal soul, whose moral agency is a gift and a responsibility before God. From this perspective, functionalism is criticized for reducing the moral agent to an earthly mechanism, ignoring the transcendent dimension of conscience (the inner “voice of God,” as Newman called it) and the need for God’s grace in moral life.

While personalist philosophers focus on the internal dimension of moral subjectivity, other contemporary philosophers emphasize the social and narrative context that functionalism overlooks. Alasdair MacIntyre, in “After Virtue” (2007) and subsequent works, argues that contemporary moral discourse has lost its footing by rejecting the classical idea of human telos, or purpose. The Aristotelian worldview (which MacIntyre calls “functional concepts of human flourishing”) treated moral virtues as qualities that enable a person to fulfill their function and achieve their purpose (eudaimonia) (MacIntyre, 2020). Enlightenment philosophy rejected this functional teleology (in the other sense of the word

"functional" – meaning function or teleological purpose) and attempted to base morality on abstract principles or feelings, leaving us with a fragmented, emotional ethic. MacIntyre's project is to restore virtue ethics based on narrative and tradition. According to this approach, a person becomes a moral agent only within the framework of a specific narrative identity and social practices. Importantly, MacIntyre criticizes the contemporary bureaucratic and managerial ethos, in which individuals are expected to divide roles (employee, citizen, etc.) and follow the functional rules of institutions without a broader moral framework. This leads to what he calls a moral partitioning: people act within the narrow function of their role, suspending their personal moral judgment (MacIntyre, 2007). An extreme case of this phenomenon was analyzed by Hannah Arendt in her book Eichmann in Jerusalem – Adolf Eichmann claimed that he was "just doing his job" (performing a function) without personal malice. MacIntyre would say that contemporary societies encourage a similar way of thinking in a milder form, which undermines true moral agency by separating actions from the narrative of a whole life and the virtues that unite them. In his book "Dependent Rational Animals" (2001), MacIntyre also emphasizes that we are vulnerable and dependent beings before we become independent practical thinkers, reminding us that moral agency is developmental and dependent on community (family, culture, etc.). A purely functional definition ("does the subject meet criteria X, Y, Z?") overlooks this temporal and social aspect; we become moral agents through education, practice, and relationships, not through the immediate fulfillment of a set of requirements. MacIntyre's critique thus suggests that functionalism is ahistorical: it treats moral agency as a static set of characteristics, whereas virtues are acquired over time within a tradition.

Charles Taylor similarly emphasizes the cultural and existential conditions of moral agency. In "Sources of the Self" (1992), Taylor argues that our moral agency is shaped by strong evaluations: qualitative distinctions between goods and virtues that we acquire from our culture and reflectively endorse. He argues that contemporary technocratic views (including some functionalist social theories) reduce moral agency to thin procedural rationality or a set of preferences, overlooking the rich moral sources that actually motivate and orient agents (Taylor, 1992, p. 92). Taylor has explicitly engaged with what he calls "naturalist" or functional explanations of morality, criticizing, for example, *Moral Foundations Theory* for offering only an evolutionary, functional account of moral intuitions (e.g., that moral instincts evolved for cooperation) (Dang, 2022). While he acknowledges the insight of such descriptions, Taylor argues that human moral life cannot be explained solely in terms of biological or social function. The foundation of Taylor's philosophical

anthropology rests on his conception of humans as “self-interpreting animals” whose self-understandings are not merely descriptions of pregiven reality but constitutive of who we are. This idea, developed particularly in his 1977 essay “Self-Interpreting Animals” (Taylor 1985a, 45-76), holds that when I interpret an attraction as “love" rather than “lust”, this changes the very nature of what I experience Taylor argues that humans exist within “moral space” – frameworks of meaning giving direction and significance to our lives. As he states in *Sources of the Self*, “doing without frameworks is utterly impossible for us... living within such strongly qualified horizons is constitutive of human agency” (Taylor, 1992, p. 27). These frameworks are “quasi-transcendental equipment that make human life possible and any moral claim understandable” (Taylor, 1992, 28). Without such frameworks providing orientation toward the good, humans would experience “terrifying identity crisis” and lose what is distinctively human in agency. The functionalist narrative, according to which morality is solely about survival or social stability, “downplays the sources of morality that are inevitable in the way individuals explain their intuitions” (Dang, 2022). In other words, even if morality serves certain functions, people do not usually act solely because of those functions. They act because they see meaning in certain goals. Taylor’s work aims to re-saturate moral discourse with these deeper meanings (such as goodness, sacredness, dignity) that a reductive functionalist approach might dismiss as epiphenomenal. In this way, Taylor accuses functionalism of emptiness: it can catalog behaviors and perhaps their evolutionary benefits, but it overlooks how it feels and what it means for a moral agent to choose the good. In this way, it risks presenting a distorted picture of moral life. Taylor argues naturalistic approaches fail through eliminative reductionism - either reducing human phenomena to mechanistic terms or eliminating them entirely, never accommodating them as they actually appear in lived experience. Naturalistic approaches cannot account for the constitutive role of self-interpretation in human life or the holistic understanding required for human meanings. Following Heidegger and Merleau-Ponty, Taylor emphasizes the "background problem" - naturalistic approaches cannot account for the taken-for-granted practices and meanings making any explicit theorizing possible.

Another dimension overlooked by functionalism is the *interpersonal* nature of moral agency, the fact that being a moral agent involves relationships with others, not just performing actions in isolation. Two thinkers who emphasize this in very different ways are P.F. Strawson presents his views on moral responsibility in his essay “Freedom and Resentment” (1963) by focusing on our attitudes in interpersonal relationships. He noted that in real life, we naturally respond to others with reactive attitudes, feelings such as

resentment, gratitude, forgiveness, which assume that we treat those others as responsible agents (Strawson 1963). Instead of starting from an abstract theory of the conditions that make responsibility possible, Strawson starts from the fact that in our interactions we treat each other as moral agents. He argues that holding someone responsible is not a matter of that person meeting some external "functional" criteria, but rather of the relationships and expectations we have as human beings in a community (Talbert, 2025). For Strawson, these reactive attitudes are "a natural expression of an essential feature of our way of life," namely our interpersonal relationships (Talbert, 2025). Crucially, moral agency is recognized through a participatory attitude: we address the other person as you, who can answer for yourself. If someone were merely a functional mechanism, we would adopt an "objective approach", treating them as an object of management or training rather than as a real responsibility. In this way, Strawson effectively argues that any conception of moral agency must preserve the reality in which we see each other as persons rather than objects. The functionalist approach carries with it the risk of shifting to a completely objective approach, judging subjects based on whether they achieve desired results, which resembles a technocratic or behaviorist view. Strawson argues that this is not how moral responsibility actually works in human life. In practice, we do not check a list of functions; we respond to others as persons with intentions. Only in exceptional cases (mental illness, coercion) we suspend reactive attitudes and treat the other person objectively, like a malfunctioning machine. Strawson's analysis therefore suggests that moral agency is intrinsically linked to being a being towards whom others can have reactive attitudes. This is a subtle but important point: being a moral agent means being a member of a moral community to which "blame" and "praise" apply, not just a place where certain actions are performed. A functionalist, ignoring the phenomenology of moral blame and outrage, may fail to see why ordinary artificial intelligence (no matter how well it behaves) does not feel part of that community. We can use artificial intelligence instrumentally and even punish it for mistakes, but we do not (at least yet) feel resentment towards it in the same way we feel it towards a human being who has committed an offense. It suggests that to truly be a participant in moral practices, something more than just function is needed (perhaps consciousness or personality, as the standard view holds).

If Strawson continues to operate within the paradigm of mutual recognition of persons, Emmanuel Levinas reverses the perspective in favor of the primacy of the Other. Levinas proposed that ethics is "the first philosophy", meaning that our fundamental relationship is ethical responsibility toward the Other, not egocentric autonomy. In works

such as “Totality and Infinity” (1961) and “Otherwise Than Being” (1974), Levinas describes the encounter with the face of the Other as a moment that calls the self to infinite responsibility. The face of the Other says “Thou shalt not kill” and imposes an obligation on me before I make any decision (Levinas, 1961, 199). This view undermines all notions of moral agency based on the capacities or functions of the self. For Levinas, moral agency begins with heteronomy, not autonomy: I am responsible before I voluntarily choose to be responsible; I am “held hostage” by the needs of the Other. “To welcome the other is to put in question my freedom” writes Levinas (Levinas, 1961, 85). In other words, true ethical action often limits one’s own unlimited freedom in obedience to the call of the Other. The functionalist view may see moral agency in terms of the subject’s capabilities: reasoning, making choices, acting according to principles. Levinas suggests the opposite: moral subjectivity consists of passivity, of being for the Other. Key elements here are the capacity for empathy, humility, and sacrifice, none of which are easily captured by functional checklists. Levinas would probably say that an approach focused on what the subject does (like calculating the right action) overlooks the asymmetrical nature of ethics. Often the most moral position is to respond to a demand that comes from outside one’s own calculations. His philosophy also suggests that something like artificial intelligence, which lacks the existential situation of encountering the sensitivity of the Other, can never be a moral agent in the full sense of the word. It can simulate care, but it does not stand in the gaze of the Other, which hurts and awakens the conscience. Therefore, Levinas, like others, considers impersonal or procedural approaches to ethics to be deeply inadequate. If to agree with him, moral agency is not about self-sufficient functions, but about being claimed by something beyond oneself: an infinite responsibility for another person.

The presented views are by no means exhaustive. They exemplify the dimensions that seem to be overlooked (or ignored) by the functionalist concept of moral agency:

Neglect of inner life and intentionality: Functionalism deliberately ignores what cannot be observed, such as conscious intentions, subjective value judgments, or a sense of duty to oneself. However, as many thinkers argue, it is precisely these internal factors that give an action moral value. Kant’s good will, Scheler’s sense of value, Taylor’s strong evaluation – all point to an internal dimension that behavior alone cannot capture. An action that looks the same on the outside can be morally different if it stems from compassion rather than compulsion. By ignoring inner life, functionalism fails to distinguish genuine moral agency from mere imitation or chance. This is reminiscent of Harry Frankfurt’s famous counterexample of the voluntarily addicted and the involuntarily addicted: both behave the

same outwardly, but only one of them acts freely. Moral agency seems to require more than just behavior; it requires authorship of the action, which is an internal phenomenon.

Questionable responsibility and free will: Functionalism lowers the bar for agency, perhaps to accommodate advanced artificial intelligence or animals, but in doing so, it obscures the question of responsibility. If we do not require that an agent be able to act differently (at least in a compatibilist sense) or understand the moral significance of its action, can we meaningfully hold it accountable? The standard view emphasizes freedom and understanding as preconditions for blame. A purely functional agent may be a system that has no "idea" of what it is doing. To call it morally responsible is probably a categorical error. Although functionalists may respond that responsibility is a pragmatic designation of certain behaviors, this seems to reduce moral blame to a tool of conditioning. However, according to P.F. Strawson, this distorts the concept of moral responsibility (Talbert, 2025). It overlooks the human significance of holding someone accountable as a member of a community. In this way, functionalism struggles with the moral significance of responsibility; it risks turning praise and blame into mere mechanisms for shaping behavior rather than responses to a person's will.

Reductionism and the loss of human dignity: Many critics view functionalism as reductionist, treating people like machines or "mere organisms". Christian philosophers in particular warn that this undermines the idea of human dignity. If moral agency is defined solely by functional outcome, then in principle a sufficiently advanced robot, or very complex artificial intelligence system could be "moral agents". Indeed, some like List speak of the responsibility of artificial intelligence. However, such an extension of the concept can undermine the value of the concept of responsibility when, for example, the "punishment" for artificial intelligence, proved causing harm, consists only in its shutdown. Extending this concept to entities without a soul or personal conscience, risks undermining the special value attributed to human agency. The Catholic personalist tradition holds that only persons (possessing intellect and will) are moral subjects in the full sense of the word, and that this status is linked to being created in the image of God. Functionalism's attempt to treat subjectivity as an impersonal set of functions is thus viewed as an affront to the uniqueness of the person. It is no coincidence that totalitarian regimes often adopt a kind of functionalism: individuals are seen as interchangeable parts of a social machine, valuable only to the extent that they perform specific roles (worker, soldier, etc.). In contrast, personalists argue that each person has a transcendental value that goes beyond their social function (Wojtyła 1979).

Blind to context and ahistorical: Functional criteria are typically static and universal (e.g., "ability to follow rules", "capacity for moral reasoning"). However, actual moral agency develops in context: historical, cultural, narrative. A small child is not yet a fully moral agent; an adult in a corrupt society may have a distorted conscience. MacIntyre and others emphasize that virtue and practical reason require the narrative unity of life and community tradition. Functionalism has difficulty accounting for this development and context. It may prematurely recognize artificial intelligence as a moral agent because it meets a certain criteria checklist, while ignoring the fact that artificial intelligence has no life history or social affiliation. In a sense, functionalism is isolating: it looks at the agent in a vacuum (inputs and outputs), while thinkers such as Levinas and MacIntyre emphasize that moral agency is fundamentally relational (to others, to tradition).

Empathy and emotions: Another weakness is the treatment of emotions. Contemporary moral psychology shows that empathy, moral emotions, and social intuitions are crucial to how we act morally A purely functional agent can coolly calculate outcomes (like a utilitarian algorithm). Many philosophers (from David Hume to Martha Nussbaum) would consider this picture incomplete, to say the least. Scheler's analysis of *Ordo Amoris* (the order of love) argues that our love and hate fundamentally determine our values and choices (Devis and Steinbock, 2024). A functionalist design could mimic some emotional responses, but it would probably not feel them. This refers to John Searle's classic argument (the "Chinese room" thought experiment) or Hubert Dreyfus's critique of artificial intelligence: syntax is not semantics, computation is not understanding. Similarly, simulated empathy is not the same as empathy. This points to a qualitative gap that functional indicators may not capture. An action taken out of genuine compassion is morally different from the same action taken out of calculated self-interest (or programming). Many argue that only a being capable of emotional concern can fully participate in moral action, because morality is not just a matter of abstract rightness, but of concern for the welfare of others. Functionalism, especially as applied to machines, struggles with this subjective aspect, and may lead to Véliz's (Véliz 2021) notion of "moral zombies*"* or Coeckelbergh's "psychopats" (Coeckelbergh 2010).

In summary, the functionalist approach, while useful for some analytical purposes, seems to undermine or ignore key elements that philosophers and theologians have identified as defining moral act. These include the conscious and free nature of the moral self, the importance of moral intention and insight into values, the inviolable dignity of each personal subject, the embedding of subjectivity in community and narrative, and affective capacities

such as empathy, which underlie moral concern. Reducing moral agency to behaviors or performance indicators is, may be considered as a gravely incomplete picture, one that might accommodate artificial agents but fails to capture the true nature of moral agency.

4.2.3. LLMs and Compressed Models of the World

The rise of powerful large language models has made them nearly synonymous with AI as a whole. In this context, examining functionalist and reductionist framings of cognitive and moral properties naturally brings into focus another theme: the supposed emergent abilities of LLMs. This view, advocated by prominent AI researchers such as OpenAI co-founder Ilya Sutskever, holds that through applying multi-step machine-learning techniques LLMs create compressed "world models". This notion is based on the observation that by scaling LLMs (by providing more data or training compute), some capabilities seem to appear suddenly. They're absent in smaller models and then cross a threshold where performance jumps from near random to useful. Classic examples including multi-step arithmetic, exam-style QA, word-sense disambiguation, and the effectiveness of chain-of-thought prompting, emphasize that these jumps are not predictable by simply extrapolating small-model trends; they often occur at specific compute scale (Wei et al., 2022). Sutskever asserts that LLMs display such properties because these systems have learned robust world models:

> "*When we train a large neural network to accurately predict the next word in lots of different texts...it is learning a world model.... This text is actually a projection of the world.... What the neural network is learning is more and more aspects of the world, of people, of the human conditions, their hopes, dreams, and motivations...the neural network learns a compressed, abstract, usable representation of that*." (Mind Cathedral, 2023)

It's worth noting that this view does not go unchallenged in the AI community. A 2022 survey of NLP researchers found that nearly half disagreed (Michael et al., 2023). Moreover, in this context the term "world model" is often used informally and lacks a rigorous definition (Mitchell, 2019). Nonetheless, as LLMs grow in significance, the possibility that they may exhibit emergent forms of "understanding" requires more closer examination. This seems crucial especially in the context of presumed AI capability to engage in moral reasoning. The question is whether, given the possibility that LLMs exhibit emergent properties, they can engage in moral reasoning solely on the basis of a "world model" learned from large text

corpora. This question is better addressed on philosophical grounds than within AI theory alone.

Early in the 20th century, Ludwig Wittgenstein highlighted the challenge of capturing moral values in language. In his work "Tractatus Logico-Philosophicus" (1922), Wittgenstein drew a sharp line between what can be expressed in words (propositions describing facts) and what can only be shown or felt. He believed that ethical values belong to the latter category. He wrote the famous sentence: "It is clear that ethics cannot be put into words. Ethics is transcendental. (Ethics and aesthetics are one)" (Wittgenstein, 1922). In other words, moral value is not an actual property of the world that language can directly describe; rather, it is something "higher" or beyond the world of facts. Any attempt to express moral absoluteness in everyday language, for Wittgenstein, would misuse language, therefore he finally calls: "Whereof one cannot speak, thereof one must be silent". This suggests that when we try to compress moral values into words, we inevitably lose something essential. A deep sense of rightness, goodness, or duty may not be fully explainable in ordinary descriptive sentences. Ludwig Wittgenstein's later philosophy fundamentally reconceptualized the relationship between language and morality. In "Philosophical Investigations" (1953), he argued that moral discourse functions as specific "language games" embedded in broader "forms of life". According to him, the meaning of moral terms does not derive from abstract definitions, but from their use in specific social practices. When we say that someone is "brave" or "cruel", we are referring to entire patterns of shared understanding and evaluation that cannot be reduced to descriptive facts alone. This Wittgenstein's concept undermines both the Platonic approach, which places moral truths in the realm of abstraction, and the computational approach, which attempts to formalize ethics using logical rules. Understanding morality requires participation in shared practices, not just the processing of statements. Considerations regarding adherence to rules (Wittgenstein,1953, §§185-243) show that moral rules gain their normative force only through practices embodied in linguistic communities. A machine learning system trained on text therefore does not assimilate raw moral facts, but traces of human language games, along with their contextual dependencies and cultural specificity. Wittgenstein suggested that understanding any language (including ethical language) requires understanding the form of life behind it: "to imagine a language means to imagine a form of life" (Wittgenstein, 1953, §19). This implies that moral language works for us because we share human forms of life (including feelings, forms of social training, etc.). If we try to transplant these words into a

completely different context (say, a machine with no human-like life form), their meaning may evaporate or change.

The implications extend beyond individual concepts to entire moral frameworks. Charles Taylor, building on Wittgenstein, argues in his book "Sources of the Self" (Taylor 1989) that language provides the "strong evaluations" that make moral experience possible. We do not first have moral intuitions that we then express in words; rather, our capacity for moral evaluation is shaped by language. Taylor points out how different sources of morality – God, nature, human reason – become accessible through historically specific vocabularies that shape what moral considerations can even arise in speakers. Wittgenstein's observations refer to long-standing debates in metaethics concerning what moral language *does*. Two broad camps can be distinguished:

- **Cognitivism**: moral statements express beliefs and aim to describe reality. If one says "charity is good," a cognitivist believes that it's a statement that can be true or false (for example, true if charity corresponds to a certain moral reality or norm). Moral realism is a common cognitive position: the view that there are moral truths or facts that can be right or wrong (perhaps independently of our opinions). According to cognitivism, a moral dispute is similar to a dispute about facts: if Alice says "X is wrong" and Bob says "X is not wrong", then they are in fact arguing about the truth of a statement (Talbert, 2025). For example, a moral realist might argue, as Richard Boyd does, that "there is one objective property that we are all talking about when we use the term 'good' in a moral context" (Tersman, 2022). Thus, language can in principle reflect moral values if those values are objective features (even if perhaps highly complex ones) of the world or of human nature.
- **Non-cognitivism** (expressivism): moral statements do not describe facts, but express the speaker's attitudes, emotions, or recommendations. This camp includes expressivists (such as A. J. Ayer and C. L. Stevenson) and prescriptivists (such as R. M. Hare). For a non-cognitivist, the statement "charity is good" is not a statement of fact, but an expression of a positive attitude toward charity (or an encouragement to engage in charitable behavior). Ayer argued that ethical statements "are not verifiable by observation" and therefore are not true statements, they are meaningless in the strict sense of stating facts, serving only to express feelings or commands. He wrote that "ethical terms are not merely expressions of feeling. They are also calculated to evoke feeling, and thus to stimulate action" (van Rojeen, 2024). Stevenson also stated that moral terms have a special "emotional significance": "a direct aura of feeling

that surrounds the word" and is associated with people's motivations (Sias, n.d.). Importantly, according to these views, moral sentences are not true; they lack truth conditions in the way that the sentence "The cat is on the chair" has truth conditions. Non-cognitivists therefore view moral language as a condensed vehicle of approval/disapproval or imperative force, rather than a vehicle of factual content.

The debate between these views highlights the limitations of linguistic representation of normativity: if non-cognitivists are right, language can never fully capture moral values as objective content. At best, it can convey our subjective attitudes or recommendations in condensed form. If cognitivists (and realists) are right, language can convey moral truths, but then the question arises as to what these truths are and how to verify them. In both cases, moral language is not a simple encoding of clear, observable facts; it is linked to human feelings, decisions, and ways of life that may not be easy to translate into another medium (e.g., a machine data structure). The philosopher J. L. Mackie is famous for adopting the position of "error theory": our moral statements are supposed to be true, but in reality, all such statements are false because there are no objective values (Mackie 1977). In this way, our moral language may systematically misrepresent reality (by claiming that there is an objective authority that does not exist), again highlighting the fundamental discrepancy between language and actual moral properties.

Another interesting angle comes from Alasdair MacIntyre, who points out the fragmented nature of contemporary moral discourse. MacIntyre, drawing on a historical perspective, argued that contemporary moral language has lost its foundations. In his book "After Virtue" (2007), MacIntyre notes that people continue to use terms such as "good", "justice", "duty", etc., but without a coherent worldview or shared tradition that originally gave these terms meaning. He illustrates this with a famous allegory of a world in which scientific knowledge has been largely destroyed and only fragments of jargon remain, it is claimed that "the language of morality is in the same state of grave disorder" in contemporary society (MacIntyre, 2007). According to MacIntyre, the Enlightenment's rejection of Aristotle's teleology (the idea that things have natural ends or purposes) caused the collapse of the coherence of moral language. We have inherited the vocabulary of virtues and duties, but without the teleological framework (e.g., the concept of human purpose or destiny) that gave that vocabulary meaning. As a result, today's moral debate often becomes endless, with no rational way to reach agreement. In MacIntyre's words, "we have lost, to a large extent if not entirely, our theoretical and practical understanding of morality" (MacIntyre, 2007, p. 2). He famously points out that in our culture, "there seems to be no

rational way to reach moral agreement" on fundamental issues; people argue using terms such as "should" and "good" as if they were objective, but these terms function rather as expressions of arbitrary preferences in a pluralistic, emotivist culture. MacIntyre clearly identifies emotivism as the dominant, hidden moral theory of the present day: "Emotivism is the doctrine that [...] all moral judgments are nothing more than expressions of preferences, attitudes, or feelings" (MacIntyre, 2007, p. 12). In his analysis, MacIntyre concludes that although few people openly declare themselves to be emotivists, our fragmented moral discourse functions as if emotivism were true. People use moral language to manipulate, express desires, or conform to identities, rather than to point to a shared truth. For example, one person's claim that "X is bad" and another's claim that "X is good" may boil down to "I really dislike X" versus "I like X," with no higher arbiter to settle the dispute, precisely because the shared criteria or teleological context that would make "good" and "bad" more than personal feelings have been lost. MacIntyre's perspective emphasizes how much moral language is dependent on context. In the Aristotelian or Thomistic traditions (to which MacIntyre later advocates a return), calling something "good" implies a connection to the purpose of that thing or to the virtues of a flourishing life. Deprived of this context, "good" becomes an empty symbol filled with subjective choice or social convention. In this way, moral values, when pressed into the language-processing engines such as LLMs without a supporting cultural narrative, can lose their content and become mere image or representation of values. This has intriguing implications for artificial intelligence: if even we humans have difficulty giving our moral terms a coherent meaning without a living tradition, how much more difficult is it to give a machine a true understanding of morality by simply loading it with our disordered language of morality.

A common theme emerges from the above perspectives: there are inherent limitations to what language itself can convey about normativity. Unlike scientific or fact-based discourse, which can often be reduced to clear data or logical rules, moral discourse resists full codification. The meaning of an ethical statement depends on factors beyond words: social practices (Wittgenstein), emotional expressions (Ayer, Stevenson), or teleological narratives (MacIntyre). Words can capture some aspects, they are our indispensable tool for discussing ethics, but they are often a poor encoding of the rich life experience and practical wisdom that are involved in a true understanding of morality. This is clearly illustrated by debates about whether moral reasoning can be reduced to algorithms or decision-making procedures. Hubert Dreyfus argued that human moral knowledge includes tacit knowledge that cannot be fully expressed through rules or language. Hubert Dreyfus, drawing on

Heidegger and Merleau-Ponty, argued that a computer cannot achieve human-like understanding because it lacks embodied presence in the world. Without a body that can actually suffer, enjoy, and engage in unprogrammed interactions, a computer cannot fully understand concepts such as harm or care. Dreyfus's critique of artificial intelligence included in his book "What Computers Still Can't Do: A Critique of Artificial Reason" (1972) argued that human intelligence (that includes moral intelligence) derives from being an embodied organism moving in real time in a physical and social environment, which cannot be replicated by purely symbolic artificial intelligence. This is consistent with Wittgenstein's view: the meaning of our language is anchored in the "stream of life". It is extremely difficult to provide artificial intelligence with the same foundation. If AI is not alive in the biological sense, then according to this view, it cannot possess true consciousness or intentionality: qualities that seem necessary for moral agency. There is a kind of practical reasoning or judgment (phronesis, to use Aristotle's term) that experts apply in context and that cannot be exhaustively described. If this is the case, any attempt to reduce moral wisdom to a list of verbal rules or codes will miss something essential: intuitive understanding of context, emotional insight, the empirical aspect of learning to distinguish good from evil.

In summary, classical philosophy of language and metaethics warn us that moral values are not easily captured in words. There is a tension between what is implicit (what our actions, attitudes, and way of life show) and what is explicit (what is said or codified). Ethics exists partly in this implicit realm. With this foundation, we can now move on to the challenge of AI: what happens when we try to teach or program ethics into an entity that (at least so far) only processes the explicit linguistic form of our values, without the full human form of life that stands behind them?

4.3. The Creative Dimension of Moral Action

The debate on Artificial Moral Agents focuses on three ethical frameworks: deontology, utilitarianism, and virtue ethics. Especially, in the context of AMAs the dominant traditions in Western moral philosophy: deontological ethics with its categorical imperatives and consequentialism with its utilitarian calculations, have long conceived moral action as fundamentally rule-governed activity. Yet a significant part of philosophical thought challenges this mechanistic understanding by presenting moral acts as fundamentally creative acts. This approach, developed most systematically by Henri Bergson, Józef Tischner, and Karol Wojtyła, alongside crucial contributions from existentialist and

phenomenological thinkers, reveals moral action as a dynamic process of self-creation, innovative response to unprecedented situations, and creative participation in the ongoing formation of moral reality itself. Rather than merely applying predetermined principles, moral agents engage in creative interpretation, value generation, and expressive action that simultaneously shapes both themselves and their moral world. This dissertation therefore introduces into the debate a previously omitted perspective: the moral act's creative dimension.

4.3.1 Creativity and the Machine

The philosophical understanding of creativity has undergone radical changes from antiquity to the present day, with each era offering a distinct framework that explained different aspects of creative phenomena.

Plato's theory of divine inspiration positioned creativity as a fundamentally supernatural phenomenon, arguing in "Ion" that poets create great works not through knowledge (technē), but through divine possession by the Muses (Plato, 534a-d). The poet becomes an "an airy thing, winged and holy" losing individual consciousness to serve as a vessel for divine expression. This concept deprives the creative individual of agency, suggesting that creativity transcends human rational capacity.

Aristotle's competing framework for *poiesis* bases creativity on rational human activity. In contrast to Plato's divine madness, Aristotle's creativity involves the conscious application of craft knowledge within a means-end relationship (Aristotle, Poetics). Aristotle understood *poiesis* as oriented to what might be: a mode of making that produces works of intellect for reflection and the cultivation of character. Over time, the artistic tradition expanded this into a robust theory of creativity. Moving between realist and imaginative poles, *poiesis* admits both the probable and the improbable, staging lively enactments that shift with form and context (Martin, 2020).

The Enlightenment revolutionized the theory of creativity through Kant's concept of genius, defined as "the innate mental predisposition (ingenium) through which nature gives the rule to art" (Kant 1790, §46). Kant identified four essential characteristics: originality as a fundamental property, exemplarity ensuring value beyond mere novelty, the inability to explain one's own processes, and limitation to the fine arts rather than science (Burnham, n.d.). His framework introduces a crucial distinction between true creativity and "original nonsense", requiring both novelty and exemplary value. The creative process involves what

Kant calls “the free play of the imagination”: spontaneous activity unrestricted by specific rules but producing “aesthetic ideas” for which no concept proves adequate.

Hume’s associative psychology offers a mechanistic explanation, according to which creativity arises from the imagination’s ability to separate and recombine ideas based on similarity, proximity, and causality. Although “nothing seems more free than the power of thought” in creating new combinations, Hume maintains that all creative materials ultimately derive from sensory experience. (Morris and Brown, 2023). This empirical limitation suggests that creativity works through recombination rather than true creation, although the combinations themselves may produce unprecedented results.

The Romantic era changed the concept of creativity, treating it as an expression of fundamental cosmic principles. Schelling equated creativity with the ultimate principle of reality, viewing artistic creation as the revelation of absolute truth through a combination of conscious technique and unconscious inspiration (Bowie, 2024). Schopenhauer, in his concept, places creative genius as temporarily transcending individual will through aesthetic contemplation, perceiving eternal ideas beyond phenomenal appearances (Sewall and Conversi, 2025). Nietzsche’s influential distinction between Dionysian and Apollonian principles reveals creativity arising from the tension between ecstatic vitality and formal restraint, whereby authentic creativity requires the rejection of established values through artistic will to power (Stoll, 2025).

Twentieth-century philosophy expanded the scope of creativity. Bergson, in his theory of *élan vital*, presents creativity as a fundamental force driving both biological evolution and human innovation, creating truly new forms rather than merely rearranging existing ones (Lawlor and Moulard-Leonard, 2022). Whitehead goes further, arguing that “creativity is the universality of universals, the ultimate metaphysical principle underlying all things without exception” (Garland, n.d.). His process philosophy presents reality as a creative synthesis in which “the many become one and are augmented by one” through creative progress toward novelty at every moment.

Heidegger’s concept of poiesis as “bringing forth” (hervorbringen) views creativity as the discovery of hidden truth rather than merely the creation of objects, whereby authentic creative technē allows beings to emerge according to their own nature (Heidegger, 1954).

Contemporary philosophy introduces the concept of rhizomatic creativity through the theoretical framework of Deleuze and Guattari (1980), in which creative connections form authentic multiplicity through non-hierarchical networks. Their concept of “lines of

flight" presents creativity as unpredictable becoming rather than planned development, with creative activities being collective rather than individual in nature.

Novelty appears to be the main criterion distinguishing true creativity from mere production, but its definition remains a subject of debate in various philosophical traditions. Margaret Boden's (1994) influential distinction between psychological creativity (P-creativity) and historical creativity (H-creativity) provides the necessary conceptual clarity. P-creativity includes ideas that are novel to the creator, regardless of whether they have appeared before, while H-creativity requires recognition as novel by the general public (Boden, 1994). This distinction is crucial for evaluating artificial intelligence systems that can achieve P creativity by generating results that are new relative to the training data but are incapable of achieving true historical innovation. Boden divides creativity into three types: combinatorial (novel combinations of known ideas), exploratory (exploration within established conceptual spaces), and transformative (deliberate transformation of the conceptual spaces themselves) (Wiggins, 2006). Transformative creativity is the most radical form, requiring the rejection of the constraints that define existing domains: ability that challenges current artificial intelligence architectures limited by training parameters (Boden, 2004).

Contemporary debates about AI creativity often rely on conflicting definitions of novelty. Computer scientists' focus on statistical novelty: results that are improbable in light of training data, differs fundamentally from philosophers' interest in conceptual novelty that changes understanding. Recent systems, such as GPT-4 and DALL-E 3, exhibit remarkable combinatorial creativity, generating results that surprise with unexpected combinations, but questions remain as to whether they achieve exploratory creativity in established fields, let alone transformative creativity that redefines conceptual spaces (Moruzzi 2025).

The "Lovelace objection", derived from Ada Lovelace's 1843 observation that "the analytical engine claims no ability whatsoever to invent anything", continues to underpin debates about machine creativity. Contemporary responses include both Bringsjord's Lovelace test, which requires AI to produce results that its designers cannot explain, and Riedl's Lovelace 2.0 test, in which human evaluators set creativity constraints that AI must meet (Bringsjord et al. 2001; Riedl 2014).

Philosophical analysis reveals novelty operating on multiple levels: syntactic (new arrangements of symbols), semantic (new meanings or concepts), and pragmatic (new uses or applications). Current AI systems excel at syntactic novelty through sophisticated recombination of patterns, but struggle with semantic novelty requiring genuine

understanding. The question of whether machines can achieve pragmatic novelty: creating tools or techniques that change human practices, remains a subject of active debate as artificial intelligence systems increasingly participate in scientific discovery and artistic production.

The epistemological dimension of novelty concerns fundamental questions about knowledge creation and whether artificial systems can generate truly new understanding, rather than merely combining existing information. Karl Popper's model of conjecture and refutation places creativity at the center of scientific progress, arguing that knowledge develops through "unjustified predictions, guesses, and temporary solutions to our problems" (Popper, 1963). This emphasis on creative hypotheses that go beyond available evidence challenges current AI systems, which operate primarily by recognizing patterns in training data rather than formulating truly novel hypotheses.

Thomas Kuhn's concept of paradigm shifts reveals scientific creativity as a complete transformation of conceptual frameworks rather than a gradual accumulation of knowledge (Kuhn, 1962). Normal science proceeds by "solving puzzles" within established paradigms, but revolutionary science requires creative leaps that establish entirely new ways of understanding phenomena. The incomparability of paradigms, their resistance to direct comparison or translation, suggests that true creativity involves not only novel combinations but also fundamental reconceptualization. Current artificial intelligence systems seem limited to normal scientific operations, optimizing within predetermined frameworks rather than generating paradigm-shifting insights.

Charles Sanders Peirce's theory of abduction provides key insights into the creative generation of knowledge. Defining abduction as "the only logical operation that introduces any new idea", Peirce distinguishes it from deduction (deriving consequences) and induction (testing hypotheses) (Peirce, n.d.). Abductive reasoning involves creative "guessing" that generates testable explanatory hypotheses, operating through what Peirce calls "economy of research" to select promising paths from among infinite possibilities. This ability to strategically select hypotheses based on intuitive judgment remains a challenge for artificial intelligence systems, which lack the experience underlying human abductive instincts. The distinction between the context of discovery and the context of justification raises fundamental questions about the epistemological capabilities of artificial intelligence. While artificial intelligence systems excel at justification tasks, verifying hypotheses through data analysis, their ability to truly discover remains a matter of debate.

Clark and Chalmers' thesis of extended consciousness shifts the understanding of cognitive boundaries, arguing that "cognitive processes do not take place exclusively in the head" (Clark and Chalmers, 1998). Their Otto thought experiment, in which an Alzheimer's patient's notebook serves the same function as biological memory, suggests that artificial intelligence systems can indeed extend human cognition, rather than merely serving as tools. This framework shifts the question from whether artificial intelligence can independently generate knowledge to how teams composed of humans and artificial intelligence generate novel understanding through distributed cognitive processes.

The epistemology of computer simulations reveals additional complexities in assessing the knowledge generated by artificial intelligence. Simulations increasingly serve as "experiments" in fields where direct experimentation proves impossible, from cosmology to climatology. However, questions remain as to whether simulation results constitute true empirical evidence or merely theoretical consequences of programmed assumptions. The framing problem, determining which information is relevant without considering all possibilities, is a fundamental challenge to the creativity of AI, as true innovation often requires precisely the kind of flexible assessment of relevance that is most difficult for computational systems (McCarthy and Hayes, 1969).

Advances in the scientific applications of artificial intelligence complicate traditional epistemological categories. Systems such as AlphaFold 3 generate predictions about protein structures that go beyond current human knowledge, while pattern recognition algorithms identify regularities invisible to human perception (Kohil and Ranganathan, 2025). These achievements suggest that artificial intelligence systems can generate what appears to be new knowledge, but philosophical questions remain as to whether this constitutes true understanding or advanced information processing without understanding.

The distinction between data processing and semantic understanding proves crucial for assessing the epistemological contribution of AI. While AI systems manipulate formal representations with remarkable finesse, humans seem to be genuinely sensitive to semantic properties as such. This difference between syntactic manipulation and semantic understanding may explain the fragility of artificial intelligence in situations beyond the scope of its training distribution, suggesting that current systems do not possess the robust understanding necessary for true knowledge creation, but only pattern extrapolation.

Applying a philosophical framework of creativity to AI systems reveals both remarkable achievements and fundamental limitations that challenge traditional theoretical boundaries. Contemporary AI systems demonstrate impressive capabilities across Boden's spectrum of

creativity, excelling particularly in combinatorial creativity through novel combinations of training data elements. Systems such as GPT-4 and Claude generate surprising linguistic combinations that exhibit statistical novelty, while DALL-E 3 and Midjourney produce visual outputs that combine elements in ways that human artists might not conceive (Moruzzi, 2025).

Exploratory creativity within established domains shows mixed results. AI systems successfully explore mathematical spaces to discover new theorems and navigate chemical possibility spaces to identify new drug compounds (Bolger, 2024). AlphaFold's predictions of protein structures represent true exploratory creativity within biochemical constraints, generating solutions that are both novel and valuable to scientific research. However, these studies remain limited to the domains in which they have been trained and struggle to transfer insights beyond the conceptual boundaries that humans navigate with ease. Transformative creativity, fundamentally changing conceptual spaces, remains the greatest challenge for AI. While humans can recognize when problems require abandoning existing frameworks, current AI architectures operate within fixed parameter spaces defined during training. The inability to modify one's own conceptual constraints suggests a fundamental limitation in achieving the paradigm-shifting creativity that characterizes major scientific and artistic breakthroughs.

Contemporary debates increasingly focus on collaboration between humans and AI, rather than on replacing humans with AI. The concept of "co-creation" recognizes that a partnership between humans and AI can produce creative results that would be impossible for either party to achieve alone. Humans provide intentionality, contextual understanding, and value judgments, while AI offers tremendous exploratory capabilities, pattern recognition, and freedom from cognitive biases. This framework for collaboration allows us to bypass unproductive debates about whether artificial intelligence is "truly" creative and focus on how the combination of humans and artificial intelligence can enhance creative capabilities (Wingström et al., 2022).

Current technical architectures impose certain limitations on AI creativity. Transformer-based models are excellent at capturing statistical regularities, but they struggle with causal reasoning and counterfactual thinking, which are essential for human creativity. The lack of embodied experience limits AI's ability to generate situational and contextual creativity that comes from real engagement with the world. Without a true understanding of goals, values, and intentions, AI systems generate outputs that may appear creative but lack the intentionality traditionally considered essential to creative acts.

Discussing the creative capabilities of machines demands calling out the event that fundamentally reshaped thinking on this subject. In March 2016, DeepMind's AlphaGo faced Lee Sedol, one of the greatest Go players of all time, in a five-game match in Seoul. In the second game, on the 37th move, AlphaGo placed a black stone in a completely unexpected place on the board. Experienced Go commentators, including professional players with many years of experience, were surprised. This move seemed to contradict centuries of human knowledge about the game. However, as the game progressed, the genius of move 37 became apparent. The stone that seemed misplaced had a subtle but powerful impact on the entire board, disrupting Lee Sedol's strategy and laying the foundation for AlphaGo's ultimate victory. The AI had not made a mistake; it had spotted an innovative and deeply strategic opportunity that went beyond human intuition and established Go theory. The debate that followed focused on the nature of AlphaGo's "creativity". Was it true ingenuity, or simply the result of deep neural networks and intensive self-play that allowed it to explore a wider range of strategies than any human could? Although the mechanisms were computational, the result was undeniably creative in its effect. It created something new, surprising, and valuable. The legacy of Move 37 continues to influence the field of AI. Every time when AI exhibits new, surprising abilities those are referred as "Move 37". In the context of the historic Go match between AlphaGo and Lee Sedol, Demis Hassabis, co-founder and CEO of DeepMind, discussed the creativity of artificial intelligence, particularly in relation to the now famous "Move 37." He proposed three levels of creativity that allow us to understand and evaluate the creative potential of artificial intelligence:

- Interpolation (basic creativity): This is the most basic form of creativity, in which an artificial intelligence system essentially averages what it has already seen. For example, if an AI is trained on a million pictures of cats, it can generate a new, average-looking picture of a cat. This is creative in the sense that the specific generated image did not exist before, but it is not a significant leap from the input data.
- Extrapolation (indirect creativity): This level of creativity involves going beyond existing data and entering new territory. Hassabis places AlphaGo's "move 37" in this category. This move was so unusual that it was initially considered a mistake by human experts. It was not a move found in the vast database of human games that AlphaGo had studied, but rather an innovative solution it discovered through its own experience gained from millions of games against itself. This demonstrated a deeper

understanding of the game and the ability to generate truly new and effective strategies.

- Invention (advanced creativity): This is the highest level of creativity, which Hassabis believes is currently unique to humans. It involves not only making a new move within an existing game but also inventing an entire game. He cites the creation of a game as elegant and beautiful as Go as an example. This level of creativity requires a deeper level of abstraction, understanding, and intention that artificial intelligence has not yet demonstrated (Sockel, 2025).

Finally, there is one more notion limiting AI's ability to create truly new things. This phenomenon is referred as non-computable problems, meaning they cannot be solved by any step-by-step set of instructions, or algorithm. The concept of non-computable problems has been introduced in 1936 by Alan Turing. In his paper: "On Computable Numbers, with an Application to the Entscheidungsproblem", Turing presented something he called the "Halting problem". Robert Marks' book "Non-Computable You: What You Do That Artificial Intelligence Never Will" (2022) calls out also the whole range of other such non-computable problems, including: Rice's Theorem, Elegant Programs (Kolmogorov Complexity), and Chaitin's Number. Marks uses the existence of these non-computable problems as a cornerstone of its argument that human abilities like genuine creativity, understanding, and consciousness are also non-algorithmic and, therefore, are things that artificial intelligence will never be able to duplicate.

4.3.2 The Creative Character of Moral Action

The foundation for the analysis of moral action as creative action is found in the work of Henri Bergson. Bergson in "The Two Sources of Morality and Religion" (1977) provides a basic framework for understanding moral action as creative rather than obligatory. Bergson identifies two fundamentally different sources of moral action: closed morality rooted in social pressure and open morality resulting from creative emotions and mystical experiences. This distinction revolutionizes ethical theory by showing how true moral action goes beyond mere rule-following to become a form of creative participation in the fundamental creative impulse of life.

Closed morality operates through what Bergson calls "infra-intellectual emotions", which generate social cohesion through duties and habits. As Vasileios Stratis explains, "social duties are imposed on individuals by society through specific rules that are narrowly defined by moral ideas, which are regarded by Bergson as products of the intellect" (Stratis

2021). This form of morality remains fundamentally limited because it “always excludes other societies” and serves primarily the survival of the tribe rather than true ethical development. Closed morality produces only superficial conformity of behavior without engaging the creative depth of the individual.

Open morality derives from creative emotions that generate new moral possibilities rather than enforcing existing social norms. Bergson’s key insight concerns the fundamental difference in how these emotions work: “in normal emotions, we first have a representation which causes the feeling (I see my friend and then I feel happy); in creative emotion, we first have the emotion which then creates representations” (Lawlor and Moulard-Leonard, 2022). This creative emotion functions like artistic inspiration, a musician creates a symphony based on emotions, which then generate musical ideas in the score. Similarly, moral creativity begins with a creative emotion, which then gives rise to new ethical insights and actions (Lawlor and Moulard-Leonard, 2022).

These creative emotions arise from what Bergson calls mystical experience, although he emphasizes that “genuine mystical experience must result in action; it cannot remain simple contemplation of God”. The “Complete mystics” achieve union with what Bergson calls “the principle of life” or “the effort of God”: understood not as a traditional deity, but as “the source of constant creativity” (Stratis 2021, 11). Through this connection with the creative principle of life, individuals transcend social conditioning and act from their “whole self” rather than under external pressure, achieving what Bergson calls “ontological freedom” (Lawlor and Moulard-Leonard, 2022).

Bergson’s understanding directly challenges Kantian ethics, revealing the categorical imperative as representing a closed morality: a “psychological error” that externalizes a unique mystical experience into a rigid moral theory in which “duty becomes harsh and inflexible”. Instead, Bergson proposes the “impulse of love” as the true moral force: a creative emotion that generates new ethical formulations through direct experience of the creative principle of life. This transforms moral action from obedience to duty into creative participation in the ongoing evolution of reality (Lawlor and Moulard-Leonard, 2022).

Drawing on Bergson’s ideas Józef Tischner argues that moral action is not just something we do; it is how we make ourselves. To name the depth from which ethical action springs, the essay borrows Heidegger’s notion of a “way of being”. It’s the spring from which deeds rise and every serious act therefore becomes an act of self-creation. The agent is both material and maker: by shaping the good in the world, a person simultaneously shapes the form of their own life. Discovering one’s ethos means spotting the “white spots” where good

is missing and the "black spots" where evil persists, then answering them with creative fidelity. This is why moral action resembles art: each deed is concrete and singular yet radiates something universal. St. Francis's gesture – giving his only clothes to the poor – cannot be deduced from a general maxim, yet it discloses a form of life oriented to love over possession. Likewise, St. Maximilian Kolbe's decision to take another prisoner's place in a death cell resists tidy rationalization; it unveils a self-made by self-gift. Such acts don't just fulfil rules; they reveal who the agent is becoming and therefore epitomize open morality. Tischner's key move is to relocate ethics from rule-following to the cultivation of a stable way of being. "Inspiration" names the inner dynamic of this self-making. In art, inspiration opens a space for what does not yet exist but could; will and feeling are drawn toward realization. Ethics works the same way, except that the artwork is the self (Tischner, 2022). Tischner develops a distinctive understanding of moral creativity through his "philosophy of drama" and his concept of the "axiological I". Tischner's contribution is to show how moral acts function as a creative self-composition through dramatic encounters with other people and values. His works show that people do not receive moral identity as a given essence but actively create it through creative responses to ethical situations and interpersonal encounters (Grabowski, 2025).

Karol Wojtyła's philosophical anthropology is probably the most systematic description of moral acts as creative acts through his personalistic understanding of human action and self-determination. In his work "The Acting Person" (1979), Wojtyła shows how moral acts are essentially creative because they involve true self-determination, in which persons simultaneously realize values and create their own moral identity. His synthesis of Thomistic metaphysics with phenomenological analysis reveals the creative structure of moral action while maintaining the objective basis for ethical evaluation. Wojtyła's main idea focuses on the principle of "operari sequitur esse": action reveals existence, but in a creative sense, not just an expressive one. As he explains: "In acting, a person not only directs himself toward a value, but also defines himself. He is not only the efficient cause of his actions, but also, in a sense, the creator of himself, especially of his moral self" (Wojtyła, 1979). This means that moral actions not only express a pre-existing moral character but actively create a person's moral identity through self-determination.

The personalistic approach highlights that freedom and self-determination are also closely related to another feature of human spiritual nature: creativity (Williams and Bengtsson, 2022). Freedom as a human characteristic allows a person to create through thoughts and actions" For Wojtyła, this creative self-determination is associated with the

choice of values that shape personal identity in a continuous process: “This particular good I am choosing has value for me according to the ‘me’ that I freely desire and choose to be” (Williams and Bengtsson, 2022). Every moral choice is therefore associated with creative self-formation, because a person not only decides what to do, but also who to become through their actions.

Wojtyła’s phenomenological analysis reveals that authentic moral actions are associated with what he calls “efficacy” – the experience of a person being the creative source of their action. As academic analysis notes, “the essence lies in recognizing the importance of human efficacy: human persons express and realize their full subjectivity through their actions.” A person experiences “I act” as “I am the effective cause” of my action, recognizing themselves as creative subjects, not just links in causal chains (Marecki 2022). This understanding bases human dignity on creative moral agency. Wojtyła builds on Kant’s categorical imperative, while incorporating it into his “personalistic principle” – persons have an inherent dignity that “surpasses all price and therefore has no equal”. Unlike Kant’s emphasis on rational duty, Wojtyła emphasizes creative freedom as the foundation of moral value: persons realize their dignity precisely through creative moral self-determination that transcends material causality (Stachewicz, 2020).

If moral actions are essentially creative actions – as Bergson, Tischner, Wojtyła argue – then the prospect of artificial moral agents faces a profound philosophical challenge. The question is not only whether machines can follow moral rules or optimize ethical outcomes, but whether they can engage in the creative self-determination, innovative response to values, and transcendent moral artistry that characterize genuine moral action. The philosophical tradition concerned with moral creativity reveals several features that seem fundamentally incompatible with current conceptions of artificial moral agents. The most immediate challenge is Bergson’s distinction between closed and open morality: machines operate entirely within closed morality, they follow programmed rules, optimize predetermined goals, and operate on what Bergson would call “intellectual” rather than creative processes. Even the most advanced machine learning systems ultimately execute deterministic or probabilistic algorithms, remaining within the realm of mechanical causality, which could be at the most equated with Bergson’s closed morality. It’s wort highlighting how contemporary artificial intelligence systems approach moral reasoning. Whether it is deontological programming (implementing specific ethical rules), consequentialist optimization (maximizing utility functions), or a virtue-based approach (imitating patterns of virtuous behavior), artificial agents essentially apply computational

procedures to moral problems. This is precisely what the creative tradition in moral philosophy considers inadequate for true moral action. As Bergson argues, authentic morality requires creative emotions that generate new moral possibilities, not calculations within existing structures. A machine executing even the most sophisticated moral algorithm remains trapped in closed morality, incapable of achieving the creative transcendence that characterizes open morality.

The problem goes beyond mere computational limitations. Tischner's concept of the "axiological I" reveals that moral identity emerges through creative response to values. The self is formed by choosing from among the values it encounters and responding to them. However, artificial agents lack this fundamental capacity for self-creation through reaction to values. Although machines can be programmed to recognize values and respond to them, they are unable to experience values as personally meaningful or use them for creative self-determination. They process value data according to pre-determined functions rather than creating themselves through creative response to values.

Wojtyła's analysis of moral action as creative self-determination probably poses the greatest challenge to artificial moral agents. His principle, according to which persons simultaneously realize values and create their moral identity through self-determined action, requires a form of agency that goes beyond material causality. Machines, no matter how advanced they are, remain entirely within the realm of material causality. Their results are determined by input data, software, and computational processes.

The phenomenological dimension that Wojtyła emphasizes: the experience of "I act" as the recognition of being the effective cause of one's own action. This dimension of experience, called "efficacy" by Wojtyła, cannot be simulated through computational processes because it involves true self-awareness and creative causality, not information processing. A machine can simulate decision-making processes, but it cannot experience itself as the creative source of its actions in the way required for moral self-determination.

One could argue that sufficiently advanced artificial systems could simulate creative moral behavior so convincingly that the distinction between genuine and simulated moral creativity would become meaningless. If a machine exhibits innovative moral reasoning, generates novel ethical solutions, and appears to engage in self-modification, does it matter whether this is "real" creativity or advanced simulation? The examined philosophical views suggest that this distinction remains crucial. Creative moral action involves not only producing novel outcomes but also transcending given conditions through genuine creative freedom. Bergson's élan vital, Tischner's axiological freedom, and Wojtyła's spiritual

transcendence point to dimensions of moral creativity that go beyond computational simulation. A machine that generates innovative moral solutions using advanced algorithms remains fundamentally different from a subject who creatively transcends their given nature through moral action. Moreover, moral creativity is linked to what phenomenologists call "lived experience" (erlebnis) the subjective first-person dimension through which values are experienced as meaningful and choices as self-determination. Without true subjectivity, machines cannot experience the creative emotions that Bergson equates with open morality, the emotional values that Tischner considers the basis of moral personality, or the self-determination that Wojtyła considers necessary for personal action.

This analysis suggests that artificial moral agents, as currently understood, are incapable of achieving true moral creativity and therefore cannot be full moral agents in the sense revealed by the tradition of moral philosophy. At best, they can function as advanced tools for moral calculation within a closed morality, implementing predetermined ethical frameworks through computational processes. This limitation has important consequences for how we should understand and use artificial agents in an ethical context.

Rather than viewing artificial systems as potential moral agents, it would be better to understand them as moral instruments: tools that can assist in human moral reasoning but are not themselves capable of creative moral action. They can help identify moral issues, assist in predicting consequences, and even generate innovative solutions within certain parameters, but they are not capable of going beyond their programming through creative self-determination or engaging in genuine moral relationships with others.

This does not diminish the importance of ensuring the ethical functioning of artificial systems or the value of research on machine ethics. However, it suggests that we should be skeptical of claims that machines can achieve true moral agency, rather than just advanced simulation of moral behavior. The creative dimension of moral action revealed by philosophers from Bergson to Wojtyła seems to require forms of transcendence, self-determination, and interpersonal interaction that remain exclusively biological or at least non-computational. Machines excel at computation but cannot grasp the normative significance intrinsic to genuine moral action. Pope Francis articulates this notion with particular clarity:

> *"No doubt, machines possess a limitlessly greater capacity than human beings for storing and correlating data, but human beings alone are capable of making sense of that data. It is not simply a matter of making machines appear more human, but of awakening humanity from the slumber induced by the illusion of omnipotence,*

based on the belief that we are completely autonomous and self-referential subjects, detached from all social bonds and forgetful of our status as creatures" (Francis, 2024).

Conclusion

The phenomenon we call artificial intelligence is deeply rooted in human efforts to build machines with human-like cognitive capabilities. This causes strong anthropomorphism tendencies in regard to the technological artifacts. The reflection on AI, in its deepest dimension, asks also about the very human nature. It raises philosophical questions about human properties and whether those can be replicated in an artificial form. Currently existing AI systems already exhibit to a great degree some human-like qualities. With the aim of building AGI (whether this is feasible or not) the things become even more complicated. Because the lack of proper terms for these new phenomena we use to describe them the vocabulary reserved so far only to animals and human beings. Therefore, we say that AI can learn, reason, hallucinate, and autonomously act in the world. The AIs capability to act makes them agents of a kind and means that results of those actions can be judged as good or wrong. This framing has profound implications. Above all, it endorses the idea of designing and building those systems in the manner that their functioning is non-harmful and beneficial for the society and the environment. This goal can be pursued using multiple means. One of them is equipping artificial agents with a capability to engage with moral deliberation. This study offers a thorough analysis of efforts to achieve this goal. Its central question is whether artificial agents can be regarded as moral agents. To address this, a multi-step methodology has been applied, combining elements of phenomenological analysis, source analysis and hermeneutics, and analytic–synthetic integration.

First, it's important to understand the complex mosaic of the AI phenomenon, coming from the fact that the term *artificial intelligence* may be confusing in the first place. It has been coined to be a "marketing" device for the newly emerging field of study in 1950s. Although attractive, already at the very beginning it invites anthropomorphizing machines. It aims at developing machines achieving human-like qualities like understanding, reasoning, learning, communicating, and acting in the world. The framing proposed by Alan Turing tried to bypass the hard question whether achieving this goal indeed requires building machines that genuinely think. To him it is enough if those technological artifacts can effectively mimic the communicative skills of humans, in the way that makes them indistinguishable from their human counterparts. This sets aside the whole history of philosophical deliberations on the human nature. Philosophers for millennia argued about the understanding and meaning of concepts like intelligence, consciousness, agency,

autonomy, thinking, morality, and creativity. The impossibility of definitive determinations at the phenomenal level, leads to a kind of lowering of the bar and, in the face of new technological artifacts, prompts us to inquiry about them in functionalist terms, allowing to speak about functional consciousness, functional morality etc. Additional complexity arises from the fact that until now there is none, one widely accepted definition of AI. That leads to the question whether they might be better framings for the phenomena at hand instead of artificial intelligence. This investigation argues that conceptual frameworks both in philosophy and computer science, proposing speaking rather about artificial agency than intelligence, offer better framing. Agency is much clearer defined. Floridi's (2023) minimalist approach require that an agent is constituted by satisfying only three basic conditions, that is, it can: (1) receive and use data from the environment, through sensors or other forms of data input; (2) take actions based on the input data, autonomously, to achieve goals, through actuators or other forms of output; and (3) improve its performance by learning from its interactions. (Floridi 2023, 12). These requirements can be satisfied by a human, animal, corporation, and a chatbot. They are also in line by the proposals on the ground of computer science, like those proposed by Russel and Norvig presented in chapter one. Speaking about artificial agents makes it also easier to relate it to the broader philosophical context. The journey from Aristotelian voluntary action through Kantian autonomy to contemporary theories of distributed and artificial agency demonstrates both remarkable continuity in core questions and reconceptualization caused by technological development. The shift from artificial intelligence to artificial agency represents more than terminological preference. It reflects an important change in defining the key terms. Luciano Floridi's thesis that AI represents "agency without intelligence" captures something essential: these systems succeed not by replicating human cognition but by achieving agential capacities through alternative means. This reframing directs attention to what artificial systems actually do (act in the world to achieve objectives), rather than speculative questions about machine consciousness or understanding. Moreover, artificial agency is not a novel phenomenon introduced by digital technology but a feature of social reality. Corporations, institutions, and simple mechanical systems have long exhibited forms of artificial agency, providing both precedents and frameworks for better understanding AI systems. The distinction between functional and phenomenological approaches to agency has important implications. If agency is essentially functional, then artificial systems might achieve genuine agency through computational processes. If agency requires subjective experience or consciousness, then current AI systems at best simulate agency without truly

possessing it. This philosophical divide cannot be resolved empirically but depends on conceptual decisions about what we mean by agency. The multiple realizability of agency suggests that artificial systems need not replicate biological mechanisms to achieve genuine agency. Just as hearts can be biological or artificial while serving the same function, agents might be biological or artificial while exhibiting the same agential capacities. However, questions about the grain of analysis and the requirements for genuine multiple realization complicate simple conclusions. Another element, the distinction between agency and personhood proves crucial for navigating ethical and legal questions about artificial systems. An entity might be a sophisticated agent without being a person deserving moral consideration, or potentially a person with limited agential capacities. This distinction helps clarify debates about AI rights and moral status.

Second, since the outcomes produced by AI systems, functioning with significant degree of autonomy, cause the impact in the social setting, the ethical dimension comes into the picture. Artificial agents act in the world and their action bring both beneficial and harmful effects. This research maintains that the question whether machines can be moral agents should be considered as a subset of broader AI ethics discussion. The fields of machine ethics and broader AI ethics share the same rationale and face the same challenges. Above all, the goal is to ensure the non-harmful AAs behavior. AI systems introduce significant concerns, namely algorithmic bias, transparency, privacy, and safety issues. They also impact the society and the natural environment. These vary from the questions about the future of work, abusive practices for exploiting human labor used to train algorithms, to the challenges related to electricity and water consumption. This highlights challenges in responding to these issues, including defining the values to be followed and deciding which ethical theory to apply. Should AI be guided by deontology, consequentialism, or virtue ethics? And regardless of the values or ethical theory selected, how should they be operationalized? This remains a significant challenge on both ethical and legal grounds. Understanding these ethical and legal dimensions is essential for discussing the morality of artificial agents, because autonomous machines are not disconnected from the world but remain part of a broader sociotechnical context. The proposed approaches by defining ethical guidelines and frameworks for trustworthy and responsible AI focus on the notion of human rights. As such they operate on high level principles to guide AI development. Legal frameworks offer some more precisely defined requirements for the AI systems. That said, both struggle with enforcing those rules in the real-world setting. Especially, with the big tech companies' approach "move fast and break things" and with their powerful lobbying

base making efforts to reduce the impact of regulations on the AI development. Additionally, with framing AI systems as tools of power, we are dealing with a kind of a "arms race" between world key players.

Third, as a part of efforts toward ensuring safe and non-harmful AI development the new field of machine ethics emerged, introducing the notion of artificial moral agents. The journey through various approaches to building moral machines – from rule-based systems following predetermined ethical principles to learning systems that develop behavioral patterns through experience – demonstrates both the ingenuity and the fundamental difficulties of this enterprise. Each approach reveals its own paradox: top-down rules fail to capture the contextual nuance of moral life, while flexible learning systems risk developing behaviors, we neither intended nor can fully predict. The hybrid approaches that attempt to combine both strategies inherit challenges from each side rather than transcending them. Questions related to those efforts multiply. The proponents of machine ethics argue that due to growing complexity of artificial agentic systems and human incapability to comprehend its functioning creation of AMAs is both inevitable and necessary. The opponents respond that the behavioral mimicry of AAs may be precisely what makes such systems dangerous: they appear moral while lacking the conscious experience, empathy, and understanding that ground genuine moral action. The Moral Turing Test and its variants crystallize this dilemma. Recent empirical studies showing that people rate machine-generated moral reasoning as superior to human reasoning, while simultaneously being able to identify it as non-human, reveal a troubling dynamic. Additionally, the machine ethics project while compelling as a concept, seems to lack efficient operationalization strategy. Also, the value alignment approach, in the context of large langue models, revealed that its implementation may be challenging and evaluation of the results even more difficult.

Fourthly, perhaps the most important insight is that questions posed by attempts at building AMAs cannot be resolved through technical means alone. The development of systems that act in morally significant ways is not merely an engineering challenge. It requires to ask questions about dimensions of morality that it tries to achieve. The analysis of functionalist approaches to artificial moral agency reveals fundamental tensions that are difficult to resolve. Although functionalism offers a path to machine ethics by focusing on observable behaviors and computational processes rather than asking about phenomenal properties, it seems omitting what might be essential for the nature of human morality. Attempts by researchers such as Crook, Corneli, and Raper to realize moral agency through machine learning techniques exemplify a broader pattern: conflating behavioral mimicry

with genuine moral capacity. When Crook and Corneli claim to map theological concepts of the soul onto functions of the basal ganglia, and then to weighted algorithms, or when Raper proposes "raising robots to be good" through developmental training analogous to raising children, they risk making radical conceptual reductions. Especially, because they introduce functionalist reductions to those dimensions that philosophers of various traditions have recognized as constitutive of moral agency. The functionalist program faces what might be called a "depth problem": at every level of analysis, moral agency seems to require something more than functional competence. Standard philosophical critiques converge on this point from many perspectives. The Aristotelian tradition emphasizes rational deliberation directed toward human development; the Thomistic view requires free will oriented toward objective good; Kantian ethics requires an autonomous rational will with inherent dignity; phenomenologists such as Scheler emphasize intuitive perception of value through feelings; personalist thinkers such as Wojtyła ground moral agency on the irreducible mystery of personality; communitarians such as MacIntyre place virtue in narrative traditions; and Levinas places ethics in the infinite responsibility evoked by the encounter with the Other. Each of these perspectives identifies dimensions of moral life that resist functionalist reduction: intentionality, consciousness, embodiment, social embeddedness, emotional reactivity, and spiritual orientation. The linguistic turn in philosophy caused by emergence of LLMs further complicates the claims of the functionalist project. If Wittgenstein is right in claiming that moral language only acquires meaning through participation in shared forms of life, then training artificial systems on text corpora cannot yield a true understanding of morality. These systems process traces of human moral discourse without access to the experiential grounding that gives that discourse meaning. The discourse surrounding large language models illustrates this clearly. Even if we accept that LLMs develop something like "compressed models of the world" through training, as Sutskever and others suggest, this actually doesn't automatically implicate that models can form the genuine moral agency. LLMs can generate very compelling arguments about various moral scenarios and dilemmas. However, compressed representation of moral discourse is not equivalent to moral understanding, just as a detailed map of a territory is not equivalent to experiencing that territory. A model can capture statistical regularities in the way people discuss ethics without understanding what makes something truly right or wrong. Current AI systems also lack generalization capabilities that would allow them to reason at the abstract level and transfer high level concepts to novel, unseen scenarios. The development of capable LLM systems so far didn't prove to be a path toward genuine AGI.

Importantly, functionalism does not take into account what P.F. Strawson described as the participatory attitude necessary for moral responsibility. Current artificial agents, no matter how sophisticated their behavior, do not exhibit this property because they lack the presence in the world of social connections that makes such attitudes valid. The functionalist approach to artificial moral agency faces a fundamental dilemma. For moral agency to be achievable for machines, it must reduce morality to behavioral functions. In doing so, however, it removes from the debate those qualities that make moral agency morally meaningful: the capacity for genuine understanding, authentic choice, emotional engagement, and meaningful responsibility. What remains is what Véliz calls "moral zombies": agents that exhibit moral behavior without possessing genuine moral life.

Fifthly, this research introduces to the debate on the morality of artificial agents new dimension: the philosophical analysis of moral acts as creative acts. Through the work of Bergson, Tischner, and Wojtyła, it has been argued that genuine moral action involves far more than following rules or calculating optimal outcomes. It requires a form of creative self-determination through which persons simultaneously shape the world and create themselves. This tradition reveals several interconnected dimensions of moral creativity that transform our understanding of morality's nature. Bergson's distinction between closed and open morality shows that authentic moral action comes from creative emotions that generate new possibilities rather than from social pressure, intellectual calculation or blind, simple rule following. Tischner's insight, built up on Bergson's notion of open morality, claims that moral action constitutes a form of self-composition. It reveals how ethical choices shape not just external outcomes but the very being of the agent. Each significant moral act becomes an act of self-creation, where the person functions as both artist and artwork. This is why moral exemplars like Francis of Assisi or Maximilian Kolbe cannot be understood through general principles alone: their actions reveal unique forms of life that emerge through creative response to values. Wojtyła's analysis of self-determination demonstrates that moral action involves a special kind of causality that transcends material determinism. This is not merely choosing between optimally calculated options but creating new possibilities through the exercise of spiritual freedom. These insights converge on a central point: moral action in its fullest sense requires creative transcendence of given conditions. The moral agents do not simply process information about values or execute predetermined decision procedures. Rather, they engage in creative response to values that simultaneously realizes those values in the world and forms their own moral identity. This creative dimension distinguishes genuine moral action from mere behavioral conformity or computational optimization. All

of this has been supported by philosophical analysis of the concept of creativity, leading to conclusion that genuine creativity might be not achievable in machines. The true creativity requires going beyond algorithmic approach and may be considered one of non-computational problems. Using the framing by Demis Hassabis: current systems lack of the highest level of creativity specific so far only to humans. AI can excel at playing Go, achieving super-human level, it can even discover unknown strategies, but it is not capable to invent a new game, that would be so elegant, simple, and sophisticated as Go.

The implications for understanding machine systems in moral contexts are significant. If moral action essentially involves creative self-determination, transcendent freedom, and lived experience of values, then systems operating through computational processes remain fundamentally limited to what Bergson calls closed morality. They can follow rules, optimize outcomes, and even generate novel solutions within predetermined parameters, but they cannot engage in the creative transcendence that characterizes open morality. This does not mean that computational systems have no role in ethical contexts. They can serve as powerful tools for moral calculation, helping identify ethical issues, predict consequences, and explore potential scenarios. The examined philosophical tradition illuminates why moral life resists complete formalization. If moral action were simply rule-following or outcome optimization, then sufficiently sophisticated systems could in principle achieve it. But if morality essentially involves creative response to values, self-formation through choice, and transcendent freedom, then it cannot be fully captured in algorithms or decision procedures.

Finally, this investigation claims that all the above leads consequentially to practical implications for how we structure moral responsibility in contexts involving computational systems. Rather than attempting to create autonomous moral agents through increasingly sophisticated programming, we should focus on developing systems that augment human moral judgment while preserving the essentially creative and personal nature of moral action. The attempts at creating artificial moral agents grounded in functionalist views risk reducing essential moral qualities. Moreover, they pose the danger of blurring the responsibility which should always be attributed only to human beings. Claims about increasing complexity of artificial intelligence systems must not serve as a “silver bullet” and excuse for “closing the responsibility gap”. AI is a technological artifact, and the challenges posed by its rapid development should be addressed not merely by better engineering but should be regarded in broader socio-technical context. As it has been presented in this dissertation the discipline of theology can contribute to this discussion thanks to its deeply humanistic, unique focus. AI is being created to mimic and replace cognitive human capabilities. But theologically

informed analysis conducted in this study argues that there are dimensions of human nature that are fundamentally irreducible. That is not to pose the debate in antagonistic terms, arguing which views better describe and capture the very nature of morality. As it has been already argued at the beginning of this thesis, when it comes to AI development, we need pluralistic, diverse debate to ensure shaping emerging technologies in the way that can possibly the best contribute to human and environmental wheel-being. This tension is perhaps better expressed by Miguel Benasayag in his book "Funzionare o esistere" (2019): *"The problem today is that, within this integrated whole, there is a desire to artificially separate the processes of functioning from the processes of existence, even going so far as to effectively negate the latter. This text does not defend the idea that one should choose one or the other of these dimensions, but rather the need to return to this complex unity. Between these two imaginary poles, there is everything".* The way in which we phrase the debate on technology development has profound significance because first we shape our technologies and thereafter they shape us.

Living in today's world of algorithms brings new challenges that call for deep reflection on what it means to be human during this *epochal change*. Pope Francis proposes compelling perspective in this regard in his encyclical "Dilexit Nos" (2024b):

"*In this age of artificial intelligence, we cannot forget that poetry and love are necessary to save our humanity. No algorithm will ever be able to capture, for example, the nostalgia that all of us feel, whatever our age, and wherever we live, when we recall how we first used a fork to seal the edges of the pies that we helped our mothers or grandmothers to make at home. It was a moment of culinary apprenticeship, somewhere between child-play and adulthood, when we first felt responsible for working and helping one another. Along with the fork, I could also mention thousands of other little things that are a precious part of everyone's life: a smile we elicited by telling a joke, a picture we sketched in the light of a window, the first game of soccer we played with a rag ball, the worms we collected in a shoebox, a flower we pressed in the pages of a book, our concern for a fledgling bird fallen from its nest, a wish we made in plucking a daisy. All these little things, ordinary in themselves yet extraordinary for us, can never be captured by algorithms. The fork, the joke, the window, the ball, the shoebox, the book, the bird, the flower: all of these live on as precious memories 'kept' deep in our heart.*" (Francis, 2024b)

References

Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., Rando, J., Nahmias, E., & Crespo, V. (2024). Attributions toward artificial agents in a modified Moral Turing Test. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-58087-7

Aif, A. (2025). *AI and Ethics: Reconsidering the Rome Call for AI Ethics*. AI And Faith. https://aiandfaith.org/insights/ai-and-ethics-reconsidering-the-rome-call-for-ai-ethics/

AlgorthmWatch. (2020). https://inventory.algorithmwatch.org

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. https://doi.org/10.1080/09528130050111428

Anderson, M. and Anderson, S.L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28 (4): 15–26.

Anderson, S. L., & Anderson, M. (2011). A prima facie duty approach to machine ethics. In *Cambridge University Press eBooks*(pp. 476–492). https://doi.org/10.1017/cbo9780511978036.032

Antiqua et nova. Note on the Relationship Between Artificial Intelligence and Human Intelligence(2025). https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html

Aquinas, T. (1948). *Summa Theologica.* Translated by Fathers of the English Dominican Province. New York: Benziger Brothers

Aristotle. (1999). *Nicomachean ethics*. Hackett Publishing Company Incorporated.

Aristotle. (2018). Physics. Hackett Publishing Company.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. Journal of Military Ethics, 9(4), 332–341. https://doi.org/10.1080/15027570.2010.536402

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, *18*(2), 103–115. https://doi.org/10.1007/s10676-016-9389-x

Asimov, I. (1950). Runaround. In *I, Robot* (The Isaac Asimov Collection ed.). Doubleday.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.08073

Barandiaran, X. E., Di Paolo, E., Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. In *Adaptive Behavior* (Vols. 17–5, pp. 367–386) https://doi.org/10.1177/1059712309343819

Bayern, S. (2017). The Implications Of Modern Business-Entity Law For The Regulation Of Autonomous Systems. *Stanford Technology Law Review*, *19*, 93–112. https://law.stanford.edu/wp-content/uploads/2017/11/19-1-4-bayern-final_0.pdf

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8.

Benanti, P. (2016). *Homo Faber. The Techno-Human Condition*. Bolonia: EDB.

Benanti, P. (2023). The urgency of an algorethics. *Discover Artificial Intelligence*, *3*(1). https://doi.org/10.1007/s44163-023-00056-6.

Benasayag, M. (2019). *Funzionare o esistere?*.

Bender, E. M., Gebru T., McMillan-Major A., and Mitchell M. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March, 610–23. https://doi.org/10.1145/3442188.3445922.

Bergson, H. (1977). *The two sources of morality and religion* (R. A. Audra & C. Brereton, Trans.). University of Notre Dame Press. (Original work published 1935).

Boden, M. A. (1994). *Dimensions of creativity*.

Boden, M. A. (2004). *The creative mind*. https://doi.org/10.4324/9780203508527

Boden, M. A. (2016). *AI: Its Nature and Future*. Oxford University Press.

Bolger, C. (2024). *Discoveries in weeks, not years: How AI and high-performance computing are speeding up scientific discovery - Source*. Source. https://news.microsoft.com/source/features/innovation/how-ai-and-hpc-are-speeding-up-scientific-discovery/

Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And How We Get There*. Random House.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*.

Bowie, A. (2024). Friedrich Wilhelm Joseph von Schelling. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2024 ed.). https://plato.stanford.edu/archives/win2024/entries/schelling/

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in Proceedings of Machine Learning Research 81:77-91.

Burnham, D. (n.d.). *Kant, Immanuel: Aesthetics | Internet Encyclopedia of Philosophy*. https://iep.utm.edu/kantaest/

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press

Bratman, M. (2014). *Shared agency: A Planning Theory of Acting Together*. Oxford University Press, USA.

Bringsjord, S., Bello, P. and Ferrucci, D., (2001), "Creativity, the Turing Test, and the (Better) Lovelace Test," *Minds and Machines*, 11: 3–27.

Bryson, J. J. (2010). Robots should be slaves. In *Natural language processing* (pp. 63–74). https://doi.org/10.1075/nlp.8.11bry

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26. https://doi.org/10.1007/s10676-018-9448-6

Callahan & Blaine. (2025). *Autonomous vehicle hits pedestrian: Who's liable?* https://www.callahan-law.com/articles-and-expert-advice/when-an-autonomous-vehicle-hits-a-pedestrian-who-is-responsible/

Calverley, D. J. (2007). Imagining a non-biological machine as a legal person. *AI & Society*, *22*(4), 523–537. https://doi.org/10.1007/s00146-007-0092-7

Claude Opus 4 and 4.1 can now end a rare subset of conversations. (n.d.). https://www.anthropic.com/research/end-subset-conversations

Clark, A. (2008). Supersizing the mind. In *Oxford University Press eBooks*. https://doi.org/10.1093/acprof:oso/9780195333213.001.0001

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. http://www.jstor.org/stable/3328150

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y

Coeckelbergh, M. (2011). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1

Coeckelbergh, M. (2013). The Moral Standing of Machines: towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, *27*(1), 61–77. https://doi.org/10.1007/s13347-013-0133-8

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Coeckelbergh, M. (2021). Three responses to anthropomorphism in social robotics: towards a critical, relational, and hermeneutic approach. *International Journal of Social Robotics*, *14*(10), 2049–2061. https://doi.org/10.1007/s12369-021-00770-0

Condon, B. S.(2018). *We are not born as fully-formed moral agents*. The Catholic Leader. https://catholicleader.com.au/people/we-are-not-born-as-fully-formed-moral-agents

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crawford, K., & Joler, V. (2018). *Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human, Data, and Planetary Resources*. AI Now Institute and Share Lab. https://anatomyof.ai

Crook, N., & Corneli, J. (2021). The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics. *Cognitive Computation and Systems*, *3*(2), 109–122. https://doi.org/10.1049/ccs2.12024

Crook, N., Nugent, S., Rolf, M., Baimel, A., & Raper, R. (2021). Computing morality: Synthetic ethical decision making and behaviour. *Cognitive Computation and Systems*, *3*(2), 79–82. https://doi.org/10.1049/ccs2.12028

Churchill, W. (1943). Cited in: Volchenkov, D. 2018. *Grammar of Complexity: From Mathematics to a Sustainable World*. https://www.worldscientific.com/doi/10.1142/9789813232501_0007

Dang, C. T. (2022). Taylor-ing Ethics: Implications of Charles Taylor's work of Retrieval on Moral Foundations Theory. *Business Ethics Quarterly*, *33*(4), 655–681. https://doi.org/10.1017/beq.2022.10

Davis, W. (2025). *Pope Leo XIV Names AI One of the Reasons for His Papal Name*. The Verge. May 10, 2025. https://www.theverge.com/news/664719/pope-leo-xiv-artificial-intelligence-concerns.

Davis, Z., & Steinbock, A. (2024). *Max Scheler*. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2024 ed.). https://plato.stanford.edu/archives/spr2024/entries/scheler/

Deign, J. (2024). *Cisco joins Rome Call as AI ethics debate widens*. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m04/cisco-joins- rome-call-as-ai-ethics-debate-widens.html

Deleuze, G., Guattari, F. (1980). *A Thousand Plateaus: Capitalism and Schizophrenia.*

Dennett, D. C. (1984). *Elbow room: The Varieties of Free Will Worth Wanting*. MIT Press.

Dennett, D. C. (1987). *The intentional stance*. Bradford Books.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

Dennett, D. C. (1996). *Darwin's dangerous idea: Evolution and the Meanings of Life*. Penguin UK.

Dennett, D. C. (2017). *From bacteria to Bach and back: The Evolution of Minds*.

Descartes, R. (1985). *The Philosophical Writings of Descartes*. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. 2 vols. Cambridge: Cambridge University Press.

Dreyfus, H. L. (1972). *What computers still can't do: A Critique of Artificial Reason*. MIT Press.

Dzidek, T. and Sikora P. (2018). „Metody". *Poznanie teologiczne*, red. T. Dzidek, Ł. Kamykowski i A. Napiórkowski, 151–174 (Teologia Fundamentalna, 5). Kraków: Uniwersytet Papieski Jana Pawła II w Krakowie. Wydawnictwo Naukowe

European Parliament.(2017). *Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics*. 2015/2103(INL)

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, *376*(2133), 20180081. https://doi.org/10.1098/rsta.2018.0081

Floridi, L. (2023). The Ethics of artificial Intelligence. In *Oxford University Press eBooks*. https://doi.org/10.1093/oso/9780198883098.001.0001

Floridi, L. (2025). AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology*, *38*(1). https://doi.org/10.1007/s13347-025-00858-9

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, *34*(1). https://doi.org/10.1007/s11023-024-09670-4

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115. https://doi.org/10.1007/bf00485230

Francis. (2024a). *LVIII World Communications Day, 2024 - Artificial Intelligence and the Wisdom of the Heart: Towards a fully Human communication*. https://www.vatican.va/content/francesco/en/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html

Francis. (2024b). *Encyclical Letter "Dilexit nos" of the Holy Father Francis on the human and divine love of the heart of Jesus Christ*. https://press.vatican.va/content/salastampa/en/bollettino/pubblico/2024/10/24/241024b.html

Frankfurt, H. G. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Fraser, H. (2024, October 31). *Deaths linked to chatbots show we must urgently revisit what counts as 'high-risk' AI*. https://doi.org/10.64628/aa.wx4n6eavr

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

Friedman, B., Jr., Kahn, P. H., Jr., Borning, A., & University of Washington. (2003). Value sensitive design and information systems. In P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations* [Book-chapter]. M.E. Sharpe, Inc.

Froese, T., & Ziemke, T. (2008). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, *173*(3–4), 466–500. https://doi.org/10.1016/j.artint.2008.12.001

Gallagher, S. (2012). *Phenomenology*. Basingstoke: Palgrave Macmillan.

Garland, W. J. (n.d.). *The Mystery of Creativity – The Whitehead Encyclopedia*. https://encyclopedia.whiteheadresearch.org/entries/thematic/metaphysics/the-mystery-of-creativity/

Gibert, M. (2022). The case for virtuous robots. *AI And Ethics*, *3*(1), 135–144. https://doi.org/10.1007/s43681-022-00185-1

Gilbert, M. (2015). *Joint commitment: How We Make the Social World*. Oxford University Press.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in Gradient-Based neural Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1312.6211

Grabowski, A. (2025). *The human drama at the heart of Józef Tischner's philosophy of drama*. Church Life Journal. https://churchlifejournal.nd.edu/articles/the-drama-at-the-heart-of-jozef-tischners-philosophy-of-drama/

Grau, C. (2006). There is no "I" in "Robot": robots and utilitarianism. *IEEE Intelligent Systems*, *21*(4), 52–55. https://doi.org/10.1109/mis.2006.81

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.14093

Guarini, M. (2011). Computational neural modeling and the Philosophy of Ethics. In *Cambridge University Press eBooks* (pp. 316–334) https://doi.org/10.1017/cbo9780511978036.023

Gunkel, D J. (2018). *Robot Rights*. Cambridge, MA: MIT Press.

Gunkel, D. J. (2023). *Person, thing, robot: A Moral and Legal Ontology for the 21st Century and Beyond*.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1606.03137

Heidegger, M. (1954). *The question concerning technology*.

High-Level Expert Group on Artificial Intelligence. (2019). *ETHICS GUIDELINES FOR TRUSTWORTHY AI*.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., . . . Perez, E. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.05566

Hume, D. (2000). *A treatise of human nature*. Oxford University Press.

Hutchins, E. (1995). *Cognition in the wild*. https://doi.org/10.7551/mitpress/1881.001.0001

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J. T., Levine, S., Dodge, J., Sakaguchi, K., Forbes, M., Hessel, J., Borchardt, J., Sorensen, T., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2025). Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00969-6

Kant, I. (1790). *Critique of judgment*. Cosimo, Inc.

Kant, I. (1998). *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood. Cambridge: Cambridge University Press.

Kant, I. (1997). *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press.

Kuhn, T. S. (1962). The structure of scientific revolutions..

Kline, R. 2011. "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence." IEEE Annals of the History of Computing 33 (4): 5–16. https://doi.org/10.1109/mahc.2010.44.

Kohli, P. (2025). Google Cloud: the platform for scientific discovery. *Google*. https://blog.google/products/google-cloud/scientific-research-tools-ai/

Korsgaard, C. M. (2009). Self-Constitution: Agency, Identity, and Integrity. Oxford: Oxford University Press.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, June 10). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*. arXiv.org. https://arxiv.org/abs/2506.08872v1

Kulisz J. (2012). *Wiara i kultura miejscem teologii fundamentalnej*. Rhetos.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., . . . Perez, E. (2023, July 17). *Measuring faithfulness in Chain-of-Thought reasoning*. arXiv.org. https://arxiv.org/abs/2307.13702

Latour, B. (2007). *Reassembling the social: An Introduction to Actor-Network-Theory*. OUP Oxford.

Lawlor, L., & Moulard-Leonard, V. (2022). Henri Bergson. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.). https://plato.stanford.edu/archives/win2022/entries/bergson/

Levin, J. (2023). Functionalism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.). https://plato.stanford.edu/archives/sum2023/entries/functionalism/

Levinas, E. (1961). *Totality and infinity: An Essay on Exteriority*.

Levinas, E. (1974). *Otherwise than being or beyond essence*.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258. https://doi.org/10.1080/00048407212341301

List, C. (2016). What is it like to be a group agent?. *Noûs*. https://doi.org/10.1111/nous.12162

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, *34*(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

List, C. (2025). Can AI systems have free will? *Synthese*, *206*(3). https://doi.org/10.1007/s11229-025-05209-x

List, C., & Pettit, P. (2011). *Group agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Locke, J. (1975). *An Essay Concerning Human Understanding*. Edited by Peter H. Nidditch. Oxford: Clarendon Press

Marks, R. J. (2022). *Non-Computable You*. Discovery Institute.

Martin, T. (2020). Poiesis. *Oxford Research Encyclopedia of Literature*. https://doi.org/10.1093/acrefore/9780190201098.013.1080

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*. Penguin UK.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McCarthy, J. & Hayes, P.J. (1969), "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in *Machine Intelligence 4*, ed. D.Michie and B.Meltzer, Edinburgh University Press, pp. 463–502.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. 2006. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955. AI Magazine, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

MacIntyre, A. C. (2001). *Dependent rational animals: Why Human Beings Need the Virtues*. Open Court Publishing Company.

MacIntyre, A. C. (2007). *After virtue: A Study in Moral Theory*.

McIntyre, K. B. (2020). Practical Reason and Teleology: MacIntyre's critique of Modern moral Philosophy. In *Palgrave studies in classical liberalism* (pp. 279–294). https://doi.org/10.1007/978-3-030-42599-9_19

Metzinger, T. (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

Mind Cathedral. (2023). *CONFERENCE JENSEN HUANG (NVIDIA) and ILYA SUTSKEVER (OPEN AI).AI Conferenza SUB ITA* [Video]. YouTube. https://www.youtube.com/watch?v=ZZ0atq2yYJw

Mitchell, M. (2019). *Artificial intelligence: A Guide for Thinking Humans*. Penguin UK.

Mitchell, T. M. (1997). *Machine Learning*. New York: Mcgraw-Hill.

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2023). *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey*. https://doi.org/10.18653/v1/2023.acl-long.903

Misselhorn, C. (2022). Artificial moral agents. In *Cambridge University Press eBooks* (pp. 31–49). https://doi.org/10.1017/9781009207898.005

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/mis.2006.80

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, *38*(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). *Levels of AGI for operationalizing progress on the path to AGI*. arXiv.org. https://arxiv.org/abs/2311.02462v4

Morris, W. E., & Brown, C. R. (2023). David Hume. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.). https://plato.stanford.edu/archives/win2023/entries/hume/

Moruzzi, C. (2025). Artificial intelligence and creativity. *Philosophy Compass*, *20*(3). https://doi.org/10.1111/phc3.70030

Mróz, M. (2024). „Sztuczna Inteligencja Jako Szansa Dla Teologii". *Teologia i Moralność* 19 (1(35)): 247–61. https://doi.org/10.14746/tim.2024.35.1.14.

Nietzsche, F. (1887). *On the Genealogy of Morals*. Translated by Walter Kaufmann and R. J. Hollingdale.

Nietzsche, F.. (1889). *Twilight of the Idols*. Translated by R. J. Hollingdale.

Nyholm, S. (2023). *This is Technology Ethics: An Introduction*. John Wiley & Sons.

Oxford English Dictionary. (2023). "*Artificial Intelligence, N. Meanings, Etymology and More | Oxford English Dictionary*. Oed.com, December. https://doi.org/10.1093//OED//3194963277.

Paglia, V. (2024). *L'algoritmo della vita*. Edizioni Piemme.

Peirce, C. S. n.d. . *Collected Papers of Charles Sanders Peirce*, edited by C. Hartshorne, P. Weiss, and A. Burks, 1931–1958, Cambridge MA: Harvard University Press.

Plato. (n.d.). *Ion*. https://topostext.org/work/545

Polger, T. W., Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press

Pope, S. J. (2024). The Roman Catholic conceptualisation of morality: Its essence and distinctive character. *Verbum Et Ecclesia*, *45*(1). https://doi.org/10.4102/ve.v45i1.2970

Popper, K. R. (1963). Conjectures and refutations: The Growth of Scientific Knowledge. Psychology Press.

Poulsen, A., Anderson, M., Anderson, S. L., Byford, B., Fossa, F., Neely, E. L., Rosas, A., & Winfield, A. (2019). *Responses to a critique of artificial moral agents*. arXiv.org. https://arxiv.org/abs/1903.07021v1

Powers, T. M. (2011). Prospects for a Kantian machine. In *Cambridge University Press eBooks* (pp. 464–475). https://doi.org/10.1017/cbo9780511978036.031

Press Release: Cisco signs the Rome Call for AI Ethics. (2024, April 24). https://press.vatican.va/content/salastampa/en/info/2024/04/24/240424a.html

Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022). A Human Rights-Based approach to responsible AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2210.02667

Przegalinska, A., & Triantoro, T. (2024). *Converging minds: The Creative Potential of Collaborative AI*.

Putnam, H. (1975). Mind, language and reality. In *Cambridge University Press* eBooks. https://doi.org/10.1017/cbo9780511625251

Raper, R. (2022). *Raising robots to be good: A practical interpretation, framework and methodology for developing moral machines* [Doctoral dissertation, Oxford Brookes University]. Oxford Brookes University. https://doi.org/10.24384/4YPZ-D514

Raper, R. (2024). Raising Robots to be Good. https://doi.org/10.1007/978-3-031-75036-6

Regulation - EU - 2024/1689 - EN - EUR-LEX. (n.d.). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Riedl, M. (2014). *The Lovelace 2.0 Test of Artificial Creativity and Intelligence*.

Robotics Openletter | Open letter to the European Commission. (n.d.). https://robotics-openletter.eu/

Rome Call for AI Ethics. (2020). https://www.romecall.org/the-call/

Runciman, D. (2023). *The handover: how we gave control of our lives to corporations, states, and AIS*.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, Stuart, and Peter Norvig. 2021. "Artificial Intelligence: A Modern Approach, Global Edition". 4th ed. London, England: Pearson Education.

Scheutz, M. (2009). *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots*.

Schwitzgebel, E. (2023). *The full rights dilemma for AI systems of debatable moral personhood*. https://journal.robonomics.science/index.php/rj/article/view/32

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. https://doi.org/10.1017/s0140525x00005756

Searle, J. R. (2010). *Making the social world: The Structure of Human Civilization*. Oxford University Press, USA.

Sewall, R.B., Conversi, L.W. (2025). *Schopenhauer and Nietzsche*. *Encyclopedia Britannica*. https://www.britannica.com/art/tragedy-literature

Sætra, H. S. (2021). Challenging the Neo-Anthropocentric relational approach to robot rights. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.744426

Sias, J. (n.d.). *Ethical Expressivism | Internet Encyclopedia of Philosophy*. https://iep.utm.edu/eth-expr/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961f

Sockel, H. (2025, February 17). The one type of "creativity" only humans can do. *Medium*. https://medium.com/blog/the-one-type-of-creativity-only-humans-can-do-eb211da3d5c0

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Spinello, R. A. (2011). Karol Wojtyla on Artificial Moral Agency and Moral Accountability. *The National Catholic Bioethics Quarterly*, *11*(3), 469–491. https://doi.org/10.5840/ncbq201111331

Spinoza, B. (1996). *Ethics*. Translated by Edwin Curley. London: Penguin Books.

Stachewicz, K. (2020). Karol Wojtyła's philosophy of freedom. *Teologia I Moralność*, *15*(1(27)), 151–162. https://doi.org/10.14746/tim.2020.27.1.10

Stenseke, J. (2021). Artificial virtuous agents: from theory to machine implementation. *AI & Society*, *38*(4), 1301–1320. https://doi.org/10.1007/s00146-021-01325-7

Stoll, T. (2025). Nietzsche's aesthetics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2025 ed.). https://plato.stanford.edu/archives/spr2025/entries/nietzsche-aesthetics/

Stratis, V. (2021). Moral duty and moral freedom in Bergson's The Two Sources of Religion and Morality. The role of the "Great Mystics" – Vasileios Stratis. *Ethical Studies*, *Vol. 6 (2), 2021*.

Strawson, P. (1963). *Freedom and resentment*. https://philarchive.org/rec/STRFAR

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics.

Talbert, M. (2025). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition ed.). https://plato.stanford.edu/archives/fall2025/entries/moral-responsibility

Taylor, C. 1985. Self-interpreting animals. In: *Philosophical Papers*. Cambridge University Press; 1985:45-76.

Taylor, C. (1992). *Sources of the self: The Making of the Modern Identity*. Cambridge University Press.

Team, O. (2025). *AI benchmarks are meaningless, AGI should lead to 10% GDP growth in developed world: Satya Nadella*. OfficeChai. https://officechai.com/ai/ai-benchmarks-are-meaningless-agi-should-lead-to-10-gdp-growth-in-developed-world-satya-nadella.

Tersman, F. (2022). Moral disagreement. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*(Fall 2022 ed.). https://plato.stanford.edu/archives/fall2022/entries/disagreement-moral/

Timnit Gebru: Ethical AI requires institutional and structural change | Stanford HAI. (n.d.). https://hai.stanford.edu/news/timnit-gebru-ethical-ai-requires-institutional-and-structural-change

Tischner, J. (2022). *Miłość i inne wydarzenia*.

Tuomela, R. (2013). *Social ontology: Collective Intentionality and Group Agents*. Oxford University Press.

Truitt, E. R. (2015). *Medieval robots: Mechanism, Magic, Nature, and Art*. University of Pennsylvania Press.

Turing, A. (1950). "Computing Machinery and Intelligence." *Mind*, vol. LIX, no. 236.

Vallor, S. (2014). Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philosophy & Technology, 28(1), 107–124. https://doi.org/10.1007/s13347-014-0156-9

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56–66. https://doi.org/10.1016/j.cogsys.2017.04.002

van Roojen, M. (2024). Moral cognitivism vs. non-cognitivism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2024 ed.). https://plato.stanford.edu/archives/sum2024/entries/moral-cognitivism/

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & Society*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wiggins, G. A. (2006). Searching for computational creativity. *New Generation Computing*, *24*(3), 209–222. https://doi.org/10.1007/bf03037332

Williams, T. D., & Bengtsson, J. O. (2022). Personalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*(Summer 2022 ed.). https://plato.stanford.edu/archives/sum2022/entries/personalism/

Wingström, R., Hautala, J., & Lundman, R. (2022). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal*, *36*(2), 177–193. https://doi.org/10.1080/10400419.2022.2107850

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. Binker North.

Wittgenstein, L. (1953). *Philosophical investigations*. United Kingdom: Wiley-Blackwell.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv (Cornell*

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now Report 2018*. AI Now Institute at New York University.

Wojtyła, K. (Pope John Paul II). 1979. *The Acting Person*. Dordrecht: D. Reidel.

Wojtyla, K. (1993). *Love and responsibility*.

Wolf, S. (1990). *Freedom within Reason*. Oxford: Oxford University Press

Van Wynsberghe, A., & Robbins, S. (2018). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, *25*(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8

Zeff, M. (2024, December 26). Microsoft and OpenAI have a financial definition of AGI: Report. *TechCrunch*. https://techcrunch.com/2024/12/26/microsoft-and-openai-have-a-financial-definition-of-agi-report/

Zuboff, S. (2018). *The age of surveillance capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.