Artificial Intelligence as a New Form of Agency (not Intelligence) and the Multiple Realisability of Agency Thesis

Luciano Floridi[1,2]

[1] Digital Ethics Center, Yale University, 85 Trumbull St., New Haven, CT 06511, USA
[2] Department of Legal Studies, University of Bologna, Via Zamboni 27/29, 40126 Bologna, Italy

Email for correspondence: luciano.floridi@yale.edu

Abstract

When interpreting Artificial Intelligence (AI) systems, we face a clear choice: either to expand our current conception of intelligence to include artificial forms of it (the *Artificial Realisability of Intelligence* or ARI thesis), or to expand our understanding of agency to encompass multiple forms, including artificial ones that do not require cognition, intelligence, intention, or mental states (the *Multiple Realisability of Agency* or MRA thesis). This article argues that scientific evidence, common sense, Ockham's razor, and an increasing body of scholarly research favour the MRA thesis over the ARI thesis. Accordingly, AI is better understood as a new form of *agency without intelligence*. By employing the Method of Abstraction, this article provides a comparative analysis of various forms of agency—natural, biological, animal social, artefactual, human, and social—to identify the defining characteristics of AI as a novel kind of agency. Reconceptualising *AI as Artificial Agency* avoids biological and anthropomorphic fallacies, improves our understanding of AI's distinct features, and provides a stronger foundation for addressing the challenges and opportunities posed by AI technologies, as well as their future development and societal impact.

Keywords: Agency Theory, Artificial Agency, Artificial Intelligence, Comparative Analysis, Method of Abstraction.

1. Introduction

The rapid development of artificial intelligence (AI) systems has reignited significant debate about the nature of machine intelligence and its relationship to biological intelligence, particularly human intelligence. I have previously argued (Floridi 2023a; Floridi and Chiriatti 2020) that this focus on intelligence—if AI is not intelligent, why not; and if it is, what kind of intelligence it is or could become—misunderstands AI's nature, scope, and potential. Instead, a more meaningful approach is interpreting AI as a new form of *Agency without Intelligence* (hereafter Artificial Agency or AA). This alternative perspective offers a clearer understanding of AI's capabilities and limitations, while avoiding conceptual pitfalls associated with anthropomorphic or biological comparisons (Floridi and Nobre 2024). In essence, we face a dilemma: either to expand our current conception of intelligence to accommodate artificial kinds of it (the *Artificial Realisability of Intelligence* or ARI thesis), or to expand our current conception of agency to include forms of agency, such as artificial ones, that lack cognition, intelligence, intention, or mental states (the *Multiple Realisability of Agency* or MRA thesis).

Before proceeding, the attentive reader will have noticed an asymmetry in the framing of the argument. This is intentional. There is little substantive debate—beyond semantic nuances—regarding the Multiple Realizability of Intelligence thesis. It is widely accepted that many animals are intelligent, and there is an uncontroversial, albeit limited and unremarkable, sense in which AI can be considered "intelligent." For instance, if intelligence is defined as the ability to perform calculations or produce coherent texts, AI qualifies in this rudimentary sense. This perspective dates back at least to Hobbes, who equated thinking with reckoning (i.e., computing). However, the real, conceptual debate concerns whether AI can be classified as intelligent in a stronger sense, equivalent to human intelligence, or even surpassing it (i.e., superhuman intelligence). This debate underpins the ARI thesis.

By contrast, those seeking a different perspective can consider the *Artificial Realisability of Agency*, which, while more specific than the MRA thesis, fits within its broader framework. I intentionally avoid framing the dilemma in this way because it is more compelling to defend the stronger version: that Artificial Agency is simply one of many forms of agency. In the remainder of this article, I argue that scientific evidence, common sense, and Ockham's razor—along with the work of several scholars (discussed in Section 3)—strongly support the MRA thesis, challenging the philosophical view that agency is exclusively human.

This shift in perspective is not merely theoretical. As AI systems become increasingly embedded in various aspects of human society, recognising their fundamental nature as unintelligent agents—rather than intelligent systems (which may also act as agents)—is critical for their effective design, development, deployment, governance, and regulation, both ethically and legally. Furthermore, understanding AI as Artificial Agency offers valuable insights into its potential and limitations, and helps develop a more suitable framework for its development and integration into human social structures and interactions.

I am aware that fully exploring the interpretation of AI as Artificial Agency would require a book-length discussion. I have done so elsewhere (Floridi 2023b), and I shall not attempt to summarise the results here. Instead, this article seeks to advance the argument for the MRA thesis by clarifying the type of agency that AI represents. In other words, it addresses the objection that unintelligent agency, as a concept, is nonsensical. Thus, the central question is not whether AI can be considered a form of agency, but rather, assuming it is, what its nature is. For readers already persuaded by this premise, further justification is unnecessary. However, for sceptics, this exploration may serve as a tantalising *modus tollens*: if one cannot convincingly define the kind of agency AI represents, it becomes difficult to argue that it possesses agency at all. The notion of Artificial Agency as merely a *quid* or a *je ne sais quoi* is highly unconvincing. This difficulty may, in turn, lend support to the ARI thesis.

The rest of the article is organised in the following sections. In section two, I briefly introduce *the Method of Abstraction* (Floridi 2008), a widely used approach in computer science that provides a rigorous framework for analysing and designing complex systems through different interfaces, referred to as levels of abstraction (LoA). In section three, I remind the reader of the current philosophical debate about the nature of Artificial Agency and its implications. After this preparatory work, the most tedious part of the article begins. It is a rather long and detailed analysis of various forms of agency. Readers who wish to save time may skip it and jump directly to section eleven. The more tenacious readers may move to section four, which analyses three essential criteria for agency. Then, section five begins the more phenomenological and taxonomic task by looking at *Natural Agency* as the most elementary form of agency, providing a foundation for understanding more complex manifestations. Section six examines *Animal Individual Agency* (*Biological Agency*), which emerges from evolutionary processes and represents a more sophisticated form displayed by living organisms. Section seven explores *Animal Social Agency* as a distinctive form of collective agency emerging from coordinated group behaviours in non-human species. Section eight investigates *Artefactual Agency*, focusing on human-made objects and systems where purpose is externally imposed by design. Section nine analyses *Human Individual Agency* as the most advanced, naturally occurring form, while section ten examines *Human Social Agency* emerging from the collective actions of individuals, groups and institutions. In section eleven, the two kinds of readers will meet again to find a comparative analysis of these different forms of agency. This section establishes a framework for understanding Artificial Agency. Sections twelve and thirteen represent the core contribution of this article, analysing *Artificial Agency* and *Social Artificial Agency* (henceforth also *Agentic AI*) as novel forms of agency distinct from both natural, biological, mechanical, and human types. Finally, section fourteen concludes by synthesising the implications of this analysis for our understanding of AI and suggesting directions for future research and development.

A final clarification. In this article, I do not attempt to define agency formally, in terms of necessary and jointly sufficient conditions. Such an endeavour is akin to attempts to define "intelligence" (Turing *docet*). Such rich and crucial concepts—think of democracy, love, friendship, etc.—are inherently complex, fuzzy, multifaceted, and context-dependent. Contrary to "triangles" and "water", they also evolve. After all, in this article, I suggest introducing a new kind of Artificial Agency. Moreover, I suspect that such concepts are, linguistically speaking, abstract nouns better understood as handy reifications of second-order properties of relations or processes, and they should be analysed in terms of an adverbial theory.[1] For example, "intelligence" should be understood as a shorthand for "behaving intelligently", "democracy" for

[1] The philosophically informed reader may recognise here the influce of Chisolm and his adverbial theory of colours (to simplify: we do not see something red, we see something redly), see (Chisholm 1957).

"governing democratically", "justice" for "acting justly", and so forth. Looking for "intelligence" must mean looking for the relations and processes that are qualified adverbially, not for something that a system has or fails to have. This is why we typically and successfully rely on *identifying criteria* to describe such concepts as clusters of overlapping features or characteristics, using a Wittgensteinian approach and terminology (Albritton 1959; Wellman 1962). This approach will suffice for this article.

2. Levels of Abstraction in Agency Analysis

Before getting to work, let me briefly introduce a methodological clarification. The Method of Abstraction (LoA) provides a fundamental framework for understanding and analysing systems—agents in this case—across different contexts. Borrowed from computer science, the method enables precise analysis of complex phenomena by defining appropriate observational frameworks. The approach becomes particularly crucial when examining agency, as the attribution and characteristics of agency can vary significantly depending on the chosen level of analysis. While a complete discussion of the LoA methodology is beyond the scope of this article, readers interested in a detailed explanation may refer to (Floridi 2008). Here, I will focus on its relevance and offer a few remarks to clarify its application.

The LoA methodology is based on the principle that complex systems can be modelled *epistemologically*—that is, *informationally* and not *metaphysically*—at multiple *levels*, which are not necessarily hierarchical (a point I will elaborate on shortly). Each level provides a *model* of different aspects of the *system*'s static and dynamic features, that is, characteristics and behaviours. This approach acknowledges that our understanding of something inherently depends on the framework through which we choose to analyse it. Effective implementation of the LoA method requires the explicit definition of observational variables or "observables" (a potentially ambiguous term that I will address shortly). They work together to form a comprehensive analytical framework that supports theoretical understanding and practical application.

To anticipate some of the points articulated later in the article, consider an ordinary thermostat system as an illustrative case of how different LoAs reveal varying aspects of agency. At the physical LoA, one may observe its mechanical components and electrical signals, which reveal the fundamental mechanisms of operation.

Feedback loops and decision algorithms become apparent at the control system level, showing the system's regulatory capabilities. At the environmental level, the focus shifts to the temperature regulation agency, indicating how the system interacts with its surroundings. At the system integration LoA, one may observe the thermostat's interactions with broader Heating, Ventilation, and Air Conditioning (HVAC) systems, emphasising its role in larger environmental control networks.

The LoA approach offers several advantages, particularly its clarity, precision, and systematic nature. In the context of this article, it reduces ambiguity in attributing agency, facilitates clear criteria for comparison, and supports systematic evaluation. By breaking down complex systems into manageable components and identifying relevant features, the LoA method enables the identification of emergent properties and supports detailed investigation. As a comparative framework, it allows meaningful cross-system comparisons and highlights similarities and differences across agency types. Importantly, it supports pluralism without falling into relativism because different LoAs can be assessed as better or worse depending on the specific purpose for which they are adopted. This alignment of purpose involves selecting abstraction levels suitable for particular analytical goals while acknowledging the limitations and trade-offs inherent to each level.

In conclusion, the following sections apply the LoA framework to analyse various forms of agency. While I aim to avoid excessive technical detail, I hope this explanation clarifies how the approach works. However, three additional clarifications are necessary for philosophically inclined readers.

First, this article employs the LoA method in a Kantian sense, focusing on the *ontology of the model* rather than the *metaphysics of the system* being modelled. In other words, the analysis is concerned with the information we have about agency as a *phenomenon*, not with agency as a *noumenon* or a *thing-in-itself*.

Second, the term "level" is used non-hierarchically in computer science. A helpful way to think of a LoA is as an *interface*. The key is maintaining a stable LoA while examining different systems, thus avoiding category mistakes or what Aristotle called a *shift in kinds* ("metábasis eis állo genos").

Finally, the term "observable" does not imply any behavioural meaning. It is a neutral term from computer science that refers to any feature of a system being considered. For instance, when modelling ghosts in Gothic novels, "gender" and "friendliness" might serve as observables. Similarly, "mental states" can qualify as observables. Confusing, I know, but it is what we have inherited from the CS department.

3. A Very Short Introduction to the Philosophical Debate about Agency

The rapid advancements in AI have prompted a fundamental reconsideration of agency, traditionally defined as the capacity for autonomous action and deliberate decision-making (Himma 2009). The boundary between human and Artificial Agency is becoming increasingly blurred as AI systems become more sophisticated and integrated into daily life—performing tasks on our behalf, instead of us, and often better than us. This challenges traditional anthropocentric frameworks (Schreiber 2024). A central tension in this debate lies between those who argue that agency is uniquely human (Fritz et al. 2020) and those advocating for broader conceptions that include non-human entities, such as artificial systems (Floridi and Sanders 2004). For clarity, I shall label these two positions, the *standard* and the *non-standard view*.

The *standard view* is still dominant. It holds that agency requires internal mental states—such as beliefs and desires—that can cause intentional actions (Swanepoel and Corks 2024; Swanepoel 2021). According to this view, AI systems may produce effects resembling human actions but lack the requisite internal states for genuine agency and morality (Sebastián 2021).

In contrast, the *non-standard view* (also known as *functionalist*, but the label is misleading; see above the clarification regarding "observable". I certainly reject the label "functionalist" as characterising my position) challenges this anthropocentric stance. It argues that agency should not be evaluated only in terms of internal states. Instead, any entity with some degree of interactivity may qualify as an agent, with more complex forms of agency including other features, such as autonomy, adaptability, *and, eventually,* internal mental states. For example, (Dung 2024) proposes a five-dimensional framework for understanding Artificial Agency encompassing goal-directedness, autonomy, efficacy, planning, and intentionality.

The debate extends beyond theoretical frameworks to practical considerations about responsibility and accountability (Wise 1998). As AI systems make increasingly consequential decisions autonomously—ranging from hiring choices to criminal risk assessments, from medical support to educational evaluations—questions of agency become entangled with issues of transparency (Andrada, Clowes, and Smart 2023), moral responsibility and accountability (Tóth et al. 2022), and human practices (Behdadi and Munthe 2020). The practical implications of Artificial Agency become pressing, particularly in scenarios involving autonomous decision-making systems, their control, and the emerging issues of distributed responsibility. Some scholars warn that attributing agency to AI systems risks creating "responsibility gaps", in which neither humans nor machines can be held appropriately accountable for harmful outcomes (Matthias 2004; Santoni de Sio and Mecacci 2021). To address this, I have argued in favour of "distributed morality", which allocates responsibility across human and artificial agents (Floridi 2013, 2016).

Others favour a more synergetic approach. For example, (List 2021) examines parallels between group agency (like corporations) and artificial intelligence, arguing that both are non-human agents that raise similar moral and regulatory challenges regarding responsibility, rights, and legal status. The analysis potentially extends to artificial systems, supporting the view that agency need not be restricted to entities possessing consciousness or intentionality in the human sense. Along similar lines, (Laukyte 2017) maintains that, if artificial agents meet the same basic conditions of agency as group agents (like corporations)—rationality, interactivity, responsibility, and personhood— they should be legally and socially recognised as having similar rights and responsibilities. Similarly, (Symons and Abumusab 2024) suggest that AI's impact on social relationships and institutions requires careful consideration beyond individual harm, arguing that AI agency should be examined as existing in degrees rather than through traditional philosophical threshold models.

Thus, the debate has moved to include more integrative approaches, recognising different degrees and types of agency. For instance, (van Lier 2023) proposes a multidimensional framework to understand how AI systems act as agents in scientific research, helping connect theoretical concepts with practical applications in AI-driven science. This is a case of a nuanced approach that allows for a comparative assessment of agency across different types of entities. Such perspectives increasingly recognise that agency exists on a *spectrum* rather than being confined to a *binary property*.

For example, (Popa 2021) contends that AI systems may exhibit genuine agency, although this agency remains fundamentally shaped by human intentions and values because human goals remain constitutive of Artificial Agency. And (Dattathrani and De' 2023) argue that "with the new generation of technologies, such as AI, the notion of agency needs to differentiate between the actions of AI from that of traditional information systems and humans" and introduce "dimensions of agency" to differentiate agencies while not privileging any kind of agent. It seems that there may be an emerging position, if not a consensus, pointing toward understanding Artificial Agency as distinct from, yet potentially complementary to, human agency. Rather than viewing Artificial Agency as diminishing human agency, scholars may be framed as *participatory* and *interactive* (Schreiber 2024). This perspective enables a more productive analysis of human-AI collaboration (Langley et al. 2017), highlighting how artificial systems can augment human capabilities while maintaining appropriate ethical frameworks for responsibility and accountability. It is the broad approach I endorse in the rest of this article.

As the debate continues to evolve alongside advances in AI, the focus is shifting toward practical governance frameworks that accommodate various types of Artificial Agency while ensuring appropriate human oversight and responsibility. Current research explores shared agency models, hybrid human-AI systems, and the integration of Artificial Agency into existing legal and social frameworks. This may, but does not have to, translate into the attribution of *legal personhood* to artificial agents (this is a related but distinct debate; see (Novelli et al. forthcoming)). However, for all this to happen through an informed and reasonable debate, it is essential to have a clear sense of *what kind of agency* we refer to when discussing Artificial Agency. The rest of the article represents my attempt to address this.

4. Three Essential Criteria for Agency

This analysis of agency begins with three fundamental *criteria* or (epistemic) *observables* that correspond to features, properties or characteristics we attribute to the identified system. These criteria help identify and qualify agency across various domains: *interactivity*, *autonomy*, and *adaptability*. Following (Floridi and Sanders 2004), we shall see that one can identify and analyse these corresponding properties at an appropriate level of abstraction, establishing a theoretical framework that supports subsequent analysis of different forms of agency.

4.1. Interactivity

Interactivity refers to an agent's capacity to engage with its environment through mutual influence. Put simply, this means the agent can act on the environment and, in turn, be acted upon by it. The nature of this interaction can vary significantly across different forms of agency, ranging from the simple physical interactions of natural agents to the multifaceted social and informational interactions exhibited by human beings. In more complex forms of agency, interactivity implies purposeful engagement with the environment, creating a dynamic relationship rather than a simple, one-way causal effect. Often, for an agent to be interactive, it must be capable of collecting data about relevant aspects of its environment and responding meaningfully to those inputs (Dennett, 1987). This bidirectional relationship creates feedback loops, which shape both the agent and its environment over time, leading to intricate patterns of interaction and adaptation. We shall see that the sophistication of an agent's interactive capabilities varies widely across different types of agents and reflects their evolutionary or designed purposes. However, one principle is clear: without some form of interaction, there can be no agency. Thus, interactivity is the foundational criterion for all forms of agency.

4.2. Autonomy

Autonomy refers to an agent's ability to initiate state changes independently of direct external causation. This characteristic does not imply total independence from environmental influences, but rather the capacity for some self-initiated and directed action within environmental constraints. The degree and nature of autonomy vary significantly across different types of agents, from zero in some natural phenomena, to the limited autonomy of simple mechanical systems, to the complex self-direction exhibited by human beings. In this case, too, autonomy should be understood as dependent on the level of abstraction at which the agent is being analysed. This contextual understanding helps avoid confusion about the nature and extent of an agent's independence, while providing a framework for comparing different forms of autonomous behaviour. We shall see that the relationship between autonomy and environmental constraints becomes significant when considering artificial agents, whose bounded autonomy must be carefully designed and constrained by human parameters.

4.3. Adaptability

Adaptability refers to the capacity of an agent to modify its behaviour based on input, such as changing physical conditions or acquiring data, information, or experience. The capacity for adaptation enables agents to maintain or improve their effectiveness in achieving their implicit or explicit goals over time. Like before, adaptability mechanisms also vary significantly across different forms of agency, reflecting the capabilities and limitations of different agents. In natural systems, adaptation occurs through physical processes and basic environmental responses. Biological systems show more sophisticated forms of adaptation through learning and behavioural modification, which are often guided by evolutionary mechanisms. Artificial systems can exhibit some programmed adaptability or, in the case of AI, advanced learning algorithms capable of complex pattern recognition and behaviour optimisation, which still fail to cope with new circumstances or situations. For example, AI models struggle to cope with novel circumstances, especially when applied to tasks outside their training parameters.

Having introduced the three essential criteria ("observables" in the useful but misleading vocabulary of the method of abstraction) that enable us to analyse different forms of agency, let us now focus on each separately.

5. Natural Agency

Natural agency represents the most elementary form of agency. It characterises non-living systems that interact with their environment. These interactions affect the environment and the agent, driven entirely by physical processes and laws. Natural agents exhibit essential characteristics that form the foundation for more complex types of agency. Their interactions operate through direct physical influence, adhering to universal principles and constants, which produce predictable patterns and environmental changes (Gell-Mann 1994). Natural agents show how complex effects may arise without conscious direction, purpose, or design. Because they may resemble more complex or sophisticated forms of agency, natural agents have led to the misattribution of higher forms of agency to non-living systems, either directly through animism or indirectly through teleological arguments for the existence of God, such as the Watchmaker argument or the Fine-Tuning argument. While the risk of this shift should be avoided (I do not add "obviously" because teleological arguments are still popular among philosophers and theologians), its presence lends support in favour of the correct recognition of some non-living systems as (some kind of) agents.

5.1. Example

Rivers provide an illustrative example of natural agency, showing how non-living systems can shape, and be shaped by, environments through consistent physical interactions. These include the erosion of riverbanks and the creation of valleys, sediment transport and deposition forming deltas, the creation of ecological niches and habitats, and influence on local climate patterns and water cycles. Such interactions emerge from the consistent application of physical laws. Rivers also exhibit systemic properties that show the complexity of natural agency. These include self-organising flow patterns, dynamic equilibrium maintenance, feedback loops with surrounding ecosystems, and long-term geological impacts (Haken 1983).

5.2. Limitations

Natural agency is fundamentally constrained by the absence of key features required for more complex forms of agency. These constraints are intuitive, but it may be worth stressing them here because they will reappear below in positive terms (as features present rather than absent). They include limitations in physical interactions, determination by physical laws, the absence of intentionality or purpose, no capacity for goal-directed behaviour, absence of information processing capabilities, lack of memory or historical learning, inability to learn from experience, and the inability to modify basic behavioural patterns. There is no *autonomy* or *adaptability* as previously described, only interactions. Similar constraints define the boundaries of natural agency while revealing its fundamental characteristics (Prigogine and Stengers 1984).

5.3. Implications

The study of natural agency reveals several essential principles that should inform our understanding of agency more broadly (Kauffman 1993), especially when discussing Artificial Agency and agentic AI. First, agency as interaction can occur without consciousness, intention, or a mental life. It is a natural phenomenon that challenges anthropocentric assumptions about its nature. Second, natural agency shows how complex effects can emerge from simple mechanisms, illustrating the power of consistent physical processes and how physical laws can produce organised behaviour without teleology or intentionality. The ability of natural systems to create and maintain persistent patterns through environmental interactions provides important insights into the fundamental nature of agency.

6. Biological Agency

Biological agency emerges through evolutionary processes and represents a more sophisticated form of agency displayed by living systems or organisms (Mayr 1997). In contrast to natural agents, biological agents exhibit degrees of autonomy and adaptability, including purposeful behaviour focused on survival and reproduction, internal state maintenance through homeostasis, environmental responsiveness and adaptation, decision-making capabilities, elements of intentionality, and the ability to learn from experience. Its mechanisms include genetic programming, metabolic regulation, sensory processing (neural processing in more complex organisms), response systems, and behavioural plasticity. These mechanisms work together to create integrated systems capable of self-maintenance and adaptation to dynamic environmental conditions. The complexity of these mechanisms varies significantly across species, reflecting their evolutionary history and ecological niches.

6.1. Example

Dogs provide a compelling example of biological agency, showing sophisticated behavioural patterns and cognitive functions that highlight the capabilities of biological agents. Their goal-directed behaviour in seeking food and shelter, social communication and interaction with humans and other animals, ability to learn from experience and training, adaptation to environmental changes, and basic problem-solving abilities (Floridi 1997) showcase the complex nature of biological agency. Dogs' cognitive functions include memory formation and recall, emotional responses, basic causal understanding, social cognition, spatial navigation, and communication skills. They illustrate how biological agents exhibit degrees of *autonomy* and *adaptability*, in contrast to natural agents, transcending simple stimulus-response patterns, while still operating within the constraints of their genetic programming and environmental conditions (Bekoff 2002).

6.2. Limitations

Despite its sophistication, biological agency operates within significant, inherent limitations shaped by evolutionary history and biological limitations (Dennett 1996). These include ecological niche restrictions, environmental dependencies and contextual constraints, instinctual drives and genetic predispositions, species-specific cognitive capabilities, physical and metabolic constraints, learning capacity boundaries, communication limitations, limited capacity for abstract reasoning, and restricted behavioural repertoires. These limitations, and the crucial fact that they remain insurmountable, reflect the evolutionary trade-offs that define the scope of biological agency while distinguishing it from simpler natural and more complex forms of agency.

6.3. Implications

The study of biological agency offers valuable insights that help bridge the gap between simple natural agency and more complex forms while highlighting the role of evolution in developing agency capabilities (Kauffman 2002). The evolution of increasingly complex agency forms, the role of environmental pressures in shaping agency, the relationship between agency and consciousness, and the development of social behaviours all emerge as key themes. Organisational principles include integrating multiple agency levels, hierarchical control systems, feedback mechanisms, and adaptive responses. The emergence of learning and adaptation in biological systems indicates how agency can develop beyond simple physical interactions while remaining grounded in material processes. It shows how complex behavioural capabilities can emerge from integrated systems of simpler mechanisms.

7. Animal Social Agency

Animal social agency represents a distinctive form of collective agency that emerges from the coordinated behaviours of groups in non-human species. It occupies a unique position between biological and human social agency (Wilson 1975; 1982), exhibiting patterns of collective action without formal or conventional organisational structures. Unlike human social agency, animal social agency emerges exclusively through evolutionary, instinctual, behavioural, and learned processes rather than conscious and deliberate teleological design. Its features include emergent leadership patterns, role specialisation, collective decision-making processes, information-sharing and learning networks, and territorial organisation. These features enable coordinated group action—such as coordinated hunting behaviours, collective defence mechanisms, and resource-sharing systems—while maintaining survival advantages, adaptive flexibility, and group cohesion, thus creating resilient social systems.

7.1. Example

Ant colonies provide a classic example of animal social agency. Colony structures include division of labour, nest construction, communication networks, resource distribution systems, foraging patterns, brood care, and collective defence mechanisms. These group behaviours emerge without centralised control yet maintain remarkable efficiency and adaptability. They show how animal social agency can exhibit greater capability than individual agency, maintaining coherence through evolved social mechanisms rather than formal institutions.

7.2. Limitations

Biological limitations shape animal social agency through species-specific communication limitations, cognitive boundaries, and group size constraints. Environmental and ecological dependencies further define the operational boundaries of animal social agency through predation pressures, resource availability, habitat requirements, and seasonal variations. These constraints create challenges and opportunities for animal social agents while shaping their development and operation.

7.3. Implications

The evolution of animal social agency shows how complex social agency can emerge from relatively simple, individual biological agency, and collective goals can be achieved without requiring the formal structures characteristic of human social agency, such as rules, norms, and conventions. Its study provides crucial insights for analysing biological and artificial forms of collective agency, revealing fundamental principles of group coordination, adaptation, and collective behaviour across different domains. This is also valuable when studying artificial social agency, from multiagent systems to Agentic AI.

8. Artefactual Agency

Artefactual agency refers to the agency exhibited by artefacts, that is, human-made objects and systems. Unlike biological or social agents, artefactual agents' purposes, operational parameters, and behaviours are explicitly defined through human intention, though they may exhibit varying degrees of autonomy and adaptability within these constraints (Simon 1996). The fundamental nature of artefactual agency emerges from the intersection of designed functionality and operational autonomy. Defining features of artefactual agency include programmed or designed behaviour patterns, specific functional purposes, mechanical or algorithmic decision-making processes, limited autonomy and adaptability within designed parameters, and external goal specification (the latter often tied to machine-learning techniques). The operational characteristics encompass predictable response patterns, defined operational boundaries, engineered feedback mechanisms, structured interaction protocols, control mechanisms, and specific maintenance requirements.

8.1. Example

Smart thermostats offer an illustrative example of artefactual agency. These systems demonstrate environmental interaction, autonomy, and adaptability within the constraints of their design. They implement features that span from autonomous temperature adjustment to occupancy detection, while incorporating learning mechanisms for user preferences, environmental monitoring of humidity, optimization of energy usage, and remote-control capabilities. Operationally, a smart thermostat processes data through an integrated chain of functions. It begins with environmental sensing of temperature, humidity, and occupancy, proceeding to pattern recognition of daily routines and seasonal changes. The system continuously performs state assessments comparing current conditions with desired ones, generates appropriate responses through HVAC system control, and monitors its performance through energy efficiency metrics. These processes enable the thermostat to interact effectively with its environment and users while maintaining comfort parameters and optimizing energy usage. Interactions are implemented through multiple channels: the thermostat exercises direct HVAC system control, makes temperature and humidity adjustments, communicates through mobile app notifications, responds to emergencies such as extreme temperature conditions, and adapts its behaviour by learning from user overrides. This implementation shows both the sophistication of modern smart home systems and their nature as designed entities serving specific purposes (climate control and energy efficiency) within constrained operational parameters.

8.2. Limitations

Artefactual agency is interactive, autonomous, and adaptable but operates within specific design constraints that fundamentally shape and limit its nature and capabilities. These constraints include predetermined response patterns, limited learning capabilities, fixed operational parameters, maintenance dependencies, and energy requirements. Functional boundaries further define the scope of artefactual agency through specific application domains, environmental limitations, performance thresholds, safety constraints, and user interface requirements. These boundaries reflect technical limitations and design choices intended to ensure reliable and safe operation while serving intended purposes. Understanding these features and limitations is crucial for ensuring the safe and effective deployment of artefactual agents, particularly in contexts where they interact with biological or human agents.

8.3. Implications

The study of artefactual agency reveals important principles about the nature of designed purposefulness and its relationship to agency. They inform both the theoretical understanding of agency and practical approaches to system design. The technological evolution of artefactual agency shows increasing system complexity, enhanced adaptive capabilities, expanded operational domains, integration of multiple subsystems, and progressive autonomy development. This evolution reveals how designed systems can incorporate increasingly sophisticated forms of agency while maintaining their fundamental nature as purposeful human creations (Norman 2013). The ways in which designed purposes are implemented through mechanical and algorithmic means, while allowing for limited forms of adaptation and learning, provide insights into the possibilities and constraints inherent in created forms of agency. Artefactual agents increasingly operate at the intersection of multiple systems and domains, requiring integration and interaction. This includes consideration of:

1. Human-Machine Interfaces: the development of effective interaction protocols that enable meaningful collaboration between human operators and artefactual agents.
2. System Integration: the coordination of multiple artefactual systems within larger operational frameworks.
3. Environmental Adaptation: the capacity for artefactual agents to operate effectively across varying conditions while maintaining reliability and safety.
4. Performance Optimisation: the ongoing refinement of operational parameters to improve effectiveness and efficiency.

These considerations highlight the growing complexity of artefactual agency and its increasing importance in modern technological systems.

9. Human Individual Agency

Human individual agency represents the most complex and advanced form of naturally occurring agency, combining biological capabilities with unique cognitive and social dimensions (Bandura 2006). This form of agency shows unprecedented levels of adaptation, autonomy, self-awareness, consciousness, intentionality, and capacity for abstract thought, setting it apart from all other natural forms of agency (Taylor 1989).

Unlike other forms of agency—such as animal or artefactual—human agency is uniquely shaped by the integration of consciousness, abstract thinking and planning, creative problem-solving, cultural and social learning, rational deliberation, emotional intelligence, moral reasoning and responsibility (including making choices and taking decisions), and metacognition (the ability to reflect on one's thoughts and actions). Other core characteristics of human agency include learning from experience, causal reasoning, complex decision-making capabilities, episodic and semantic memory, future scenario simulation, long-term goal setting and pursuit, cultural transmission, sophisticated symbolic thought and language use, and complex tool creation and use. The integration of temporal and experiential factors allows human agents to learn from history, develop complex personal identities (identity construction, narrative self-understanding) and maintain biographical coherence (a sense of self and purpose). Present moment awareness enables sophisticated real-time decision-making, planning, and action. Future scenario planning allows complex strategic thinking and goal pursuit. This temporal integration creates a unique form of agency that operates across multiple time scales while maintaining personal, social, and narrative coherence. It also supports the development of cultural knowledge and social institutions, enabling forms of collective agency (see below) that transcend individual capabilities. Here, humans are the only example of such agency, so we can move directly to the section on limitations.

9.1. Limitations

Human individual agency, while describing our most familiar form of agency, does not come without limitations or "room for improvement," as they say. Despite its sophistication, human agency operates within significant constraints. Cognitive limitations include information processing boundaries, attention capacity constraints, memory limitations, decision-making biases, and emotional influences (Kahneman 2011). These limitations shape the operation of human agency while revealing important insights about its nature and capabilities. Biological, environmental, social, and cultural constraints further shape human agency through physical limitations, biological needs, temporal constraints, resource dependencies, environmental influences, social pressures, and cultural conditioning (Giddens 1984). Such limitations may also present opportunities when looking at human-artificial agency integration (a topic I shall leave to future research).

9.2. Implications

Perhaps the most important implication of human agency in this context concerns its moral dimensions as a unique characteristic. Ethical reasoning, value systems development, moral responsibility, understanding of rights, duties, and obligations, and participation in the social contract create a form of agency that incorporates moral consideration into decision-making and action (Frankfurt 1971). This moral capability enables humans to evaluate actions not only in terms of immediate consequences but also through ethical frameworks, value systems, long-term planning, and hypothetical and counterfactual reasoning. As we shall see in the next section, the capacity for moral judgment and responsibility enables the development of complex social structures and cultural systems that extend human agency beyond individual capabilities. Other biological, sentient agents may have rights, but only human agents have duties.

10. Human Social Agency

Human social agency refers to the collective capacity of human agents to act intentionally and cooperatively within structured social systems based on complex cultural, linguistic, and institutional frameworks (Weick 1995). This form of agency operates through distributed decision-making processes, which integrate diverse inputs and considerations, and formal organizational structures, creating capabilities (e.g., learning) that surpass the sum of individual agents' contributions. Thus, human social agency can address challenges and pursue opportunities beyond individual capacities, while maintaining institutional coherence and continuity. Other structural features of social human agency include coordinated action across multiple individuals and even generations, collective intentionality, institutional memory systems, formalized procedures and rules, emergent organizational behaviours, and hierarchical control mechanisms (Scott 2013). Operational characteristics encompass coordinated action patterns, information-sharing networks, resource allocation systems, power distribution structures, control and conflict resolution mechanisms, and cultural norm maintenance. All these characteristics create resilient systems capable of responding to environmental changes while maintaining organizational coherence and effectiveness. They enable human social agency to operate at multiple scales, from small communities and organisations to global systems. This capacity for large-scale coordination makes human social agency a unique, transformative form of collective action.

10.1. Example

Modern corporations exemplify social agency through their organizational structures and collective behaviours (March and Simon 1993). Not accidentally, they are among the agents considered "legal persons" to make room for their kind of agency. Their collective behaviours manifest through market adaptation, strategic planning, innovation processes, cultural development, and stakeholder management.

10.2. Limitations

Internal constraints shape social agency through coordination and control challenges, communication barriers, competing interests, decision-making inefficiencies, cultural resistance, cognitive and informational biases (groupthink, confirmation bias, and misinformation) and political dynamics (Mintzberg 2009). External constraints further define the operational boundaries of social agency through institutional inertia, legal frameworks, market conditions, social expectations, environmental factors, resource limitations, and technological capabilities (Powell and DiMaggio 2012). These constraints create challenges and opportunities for social agents while shaping their development and operation. They should be studied better to inform the design of future kinds of Agentic AI (see below).

10.3. Implications

The study of human social agency reveals important insights about emergent properties in collective systems and system dynamics. These include governance systems that ensure accountability and performance metrics that guide development (North 1990), as well as collective intelligence that emerges from coordinated individual contributions, organisational learning that transcends individual knowledge acquisition, cultural evolution through institutional processes, institutional memory that preserves and transmits knowledge, risk distribution through organisational structures, and resource mobilisation, and social innovation through collaborative processes (Luhmann 1995).

11. AI as a New Form of Agency

The preceding sections have analysed distinct forms of agency—natural, biological, animal (individual and social), artefactual, and human (individual and social). Before discussing Artificial Agency, let us take stock of the comparative analysis conducted so far.

The previous sections have highlighted fundamental commonalities and crucial distinctions that help clarify the nature of agency and establish a conceptual foundation for understanding Artificial Agency specifically. Several structural patterns have emerged. *Natural agency* represents the most basic form, defined by *interaction* with the environment through direct physical processes, without purposive behaviour. *Biological agency* introduces *autonomy* and *adaptability*, while *animal social agency* shows emergent collective behaviours. *Human agency* adds layers of *purposefulness* (intentionality and teleology), *conscious deliberation*, and *cultural meaning-making*. This progression from simpler to more complex forms of agency does not imply a strict linear hierarchy but rather a branching pattern, with each form of agency exhibiting unique characteristics advantageous within specific contexts. For instance, while human agency shows extraordinary flexibility and creative potential, we shall see that *Artificial Agency* achieves superior performance in narrow, well-defined domains through its consistent and precise operations.

The operational mechanisms underlying different forms of agency also reveal important distinctions that shape their respective capabilities and limitations. Natural agency operates through physical laws and constants, producing consistent patterns but lacking autonomy and adaptability. Biological agency, whether individual or social, functions through genetic programming, sensory processing, and behavioural regulation, enabling autonomous and adaptive behaviour that is much less predictable, even within species-specific boundaries. Human agency uniquely combines biological mechanisms with cultural and symbolic processing, resulting in unmatched flexibility and creativity. In contrast, we shall see that Artificial Agency, operating through computational and data processing, excels at specific tasks while lacking the intelligence, intentionality, and self-determination we encounter in biological systems.

Environmental interaction also varies significantly across agency types, reflecting fundamental differences in how agents engage with and influence their surroundings. Human agency, in particular, stands apart in its unbounded capacity to modify and create new physical and symbolic environments. Most significantly, this ability to shape environmental conditions introduces a unique dynamic that influences the operation of other forms of agency. Indeed, *human agency is the only agency that shapes all other agencies*.

Mechanisms of adaptation and learning provide another point of comparison. Natural agents exhibit no learning capacity, relying instead on fixed physical properties. Biological agents develop through species-specific constraints, while artificial agents learn through algorithmic optimisation. Human agency, in turn, combines individual learning with cultural transmission, creating cumulative knowledge systems that transcend individual capabilities. These distinctions highlight the unique nature of Artificial Agency as a new addition to the taxonomy of agency, blending aspects of artefactual and biological agency while introducing novel properties.

The pain is over. The comparative analysis is now complete, and the taxonomy has been established. We can now focus on Artificial Agency as a distinct and novel form of agency.

12. Artificial Agency

Artificial Agency (AA) represents a novel form of agency emerging from the interplay of programmed objectives and learned behaviours. At its core, AA is a computational, goal-driven form of agency defined by human purposes. This *goal-directedness* marks a fundamental departure from biological *purposefulness*, mechanical *determinism*, and human *intentionality*. However, its distinctive properties incorporate familiar elements in unique configurations, making AA both a continuation of and a departure from established forms of agency. While natural agents operate through fixed behavioural patterns, artificial agents can modify their behaviour goal-directedly. Unlike biological agents, they lack genuine intentionality and evolutionary development but compensate with rapid, domain-specific, adaptation capabilities. They differ from social agents (see section thirteen) in their inability to form emotional bonds, yet excel in coordinated information processing. Their sophistication surpasses traditional artefactual agency through advanced learning capabilities. At the same time, their lack of consciousness, intelligence, and mental states and their inability to transcend predefined objectives through self-determination distinguish them from human agency, even as they potentially exceed specific human capabilities.

The data-driven adaptability of artificial agents emerges through statistical learning and comprehensive pattern recognition *across diverse domains*. The italics are meant to stress the more open nature of this learning, compared to more basic forms of machine learning (see the example of the thermostat) constrained by the data set on which they are trained and the circumstances in which they operate. This learning process enables the rapid processing of vast quantities of data through programmatic pathways, rather than biological ones, creating a unique form of information acquisition and behavioural adaptation. Other operational capabilities encompass sophisticated parallel processing mechanisms that make multitasking across diverse domains possible. Artificial agents can maintain continuous operation without metabolic limitations, while their distributed functionality through networked systems enables unprecedented scalability. Integrating quantifiable uncertainty in decision-making processes adds a layer of sophistication to their operational framework. The distinctive properties of Artificial Agency manifest through the seamless integration of design and emergence, where programmed rules merge with learned patterns to create adaptive behaviours.[2] This integration enables multi-scale operations that combine local and global optimization strategies, resulting in reproducible decision-making patterns that evolve through scientific and technological innovation.

12.1. Example

Large Language Models (LLMs), such as OpenAI's GPT series, provide a compelling example of Artificial Agency in practice. They consistently align with programmed objectives while developing impressive strategies to interact with and address user queries. This goal-directed behaviour illustrates the integration of fixed parameters with adaptive response mechanisms. The system shows remarkable adaptive learning

[2] A comparable phenomenon can be observed in cellular automata, where simple rule-based systems generate emergent patterns and behaviors, demonstrating how complexity can arise from straightforward, deterministic processes.

by combining extensive pre-trained knowledge with a contextual understanding of ongoing interactions. LLMs' distributed processing capabilities enable them to handle multiple complex tasks simultaneously while maintaining internal consistency across responses. This exemplifies the sophisticated integration of rule-based operation with learning-based adaptation. The multi-scale operational capacity becomes evident in their ability to process information at both specific and broader levels, generating responses that range from specific technical details to broad theoretical frameworks. They maintain reproducible patterns throughout these operations while preserving the flexibility to adapt to unique contextual requirements.

12.2. Limitations

Despite its sophisticated capabilities, Artificial Agency is fundamentally constrained by its design and operational framework. The bounded autonomy of AI systems restricts their operations within programmed parameters, creating a clear ceiling for independent action. Their heavy dependence on training data and predefined objectives shapes their agency in ways fundamentally different from biological systems. The absence of consciousness, intelligence, and understanding creates an unbridgeable gap between artificial and human agency. This limitation extends to the inability to generate truly original, chosen, or preferred purposes or goals (e.g., the ability to choose to choose), confining artificial agents to operate within pre-established frameworks. The boundary of understanding remains firmly at pattern recognition and matching, without crossing into genuine comprehension.

Artificial agents function as purpose-bounded computational agents, operating effectively within their defined parameters but unable to transcend them in ways biological agents routinely achieve. This fundamental limitation shapes both their capabilities and their potential applications. It remains a category shift in agency for this kind of technology and its future implementations. Such systems could also be described as a *syntactic form of agency*.

12.3. Implications

The emergence of AI as a new form of agency has profound implications for this technology's development, deployment, and governance. At least five essential characteristics are crucial: its informational nature in processing and pattern recognition, its instrumental aspect in serving specified purposes, its non-conscious functioning without awareness, intelligence, or emotions, its distributed operational capability across systems, and its reproducible nature ensuring consistency across implementations.

Development considerations must align with these technical characteristics while addressing emerging challenges. The focus extends beyond merely enhancing computational capabilities within ethical boundaries to understanding how Artificial Agency will evolve, also in relation and interactions with other forms of agency. As AI systems become increasingly sophisticated, their agency characteristics will likely develop along paths distinct from biological or human agency patterns. This evolution necessitates prioritising reliability and reproducibility[3] while maintaining realistic expectations about the potential for artificial systems.

Integration with human social structures presents unique challenges that require careful consideration of agency boundaries and constraints (Pearl & Mackenzie 2023). Deploying artificial agents demands a clear distinction between artificial and human agency in practical applications, leading to the appropriate allocation of control and responsibilities. This integration must account for artificial agents' complementary capabilities and inherent limitations. Governance frameworks must also evolve beyond traditional regulatory approaches designed for animal and human agents. These frameworks must account for artificial agents' distributed and scalable nature while establishing ethical guidelines that recognize their distinct operational nature.

The reconceptualisation of AI as Artificial Agency means that its development should acknowledge and work within its distinct agency type rather than attempting to replicate human (or even just animal) intelligence. This will help avoid anthropomorphic fallacies while maintaining realistic expectations about its capabilities and limitations. The future of Artificial Agency lies not in attempting to

[3] The concept is under pressure since biological reproducibility, understood as the process by which living organisms produce offspring to perpetuate their species, is joined by artificial reproducibility, here understood as the ability of artificial systems, such as AI or robots, to create copies or improved new versions of themselves or other artificial systems.

transcend its fundamental nature but in optimising its unique characteristics for beneficial applications (Floridi 2023).

13. Social Artificial Agency or Agentic AI

Social Artificial Agency (Agentic AI) represents an emergent form of collective agency, where autonomous, coordinated AI systems interact to achieve complex goals with minimal human oversight (Shavit et al. 2023). Unlike traditional multiagent systems (MAS), which rely on predefined protocols for distributed problem-solving, Agentic AI integrates advanced capabilities such as dynamic reasoning, real-time adaptability, and multi-scale operational coordination. These systems actively intervene in environments, functioning as agents of action and influence.

Agentic AI builds on the foundational principles of Artificial Agency discussed in the previous section, while introducing new elements of collective functionality. Through the integration of large language models (LLMs), advanced pattern recognition, reinforcement learning, and planning algorithms, Agentic AI exhibits coordinated behaviours that resemble emergent collective intelligence. This development marks a significant shift in the study of agency, as it challenges traditional boundaries between individual and collective action and between human and artificial agency.

Agentic AI systems operate through networks of interconnected modules, each capable of autonomous decision-making while maintaining coordination across the system. This decentralised structure allows for scalable and resilient operation. At the core of Agentic AI is a reasoning engine, often powered by LLMs, which enables real-time decision-making and adaptability. These systems can process complex inputs, evaluate alternatives, and generate context-sensitive responses. Perception modules interpret and process environmental inputs, while action modules execute decisions in physical or digital environments. This integration facilitates seamless interaction with dynamically changing contexts. Agentic AI operates across multiple timeframes, from microsecond interactions to long-term scenario planning. This temporal adaptability allows for rapid responses to immediate conditions and the coordination of strategic actions over extended periods. This capability enables instantaneous learning distribution across networks, contrasting sharply with the generational evolution observed in biological systems. Through digital protocols, these systems achieve immediate communication across vast distances, processing enormous amounts of data at speeds that surpass biological limitations. The result is a unique hybrid that combines deterministic programming with emergent adaptation. While Agentic AI systems are designed to follow programmed objectives, their interactions often produce emergent behaviours not explicitly anticipated by their designers. These behaviours arise from the complexity of their interactions and the adaptability of their learning processes.

Unlike biological social systems, Agentic AI lacks consciousness or intentionality. However, it exhibits forms of collective intelligence through coordinated information processing, resource allocation, and goal-directed action. These fundamental differences shape both its capabilities and limitations.

Finally, Agentic AI has superior capabilities in several crucial areas compared to MAS. Its enhanced dynamic reasoning through LLMs enables unprecedented adaptability to unforeseen challenges. The system exhibits remarkable scalability in processing vast datasets, while its learning capabilities and integrated decision-making continue to set new AI.

13.1. Example

Agentic AI is a relatively new phenomenon if understood in technical and strict terms (many old applications are often merely rebranded). Instead of providing a practical example of an application, let me refer to the hypothetical textbook example of an Agentic AI travel planner. Such a system would autonomously plan flight bookings, hotel reservations, and itinerary adjustments based on user preferences and real-time disruptions. This hypothetical example highlights the potential benefits of Agentic AI, including enhanced efficiency, convenience, and scalability. However, it also underscores the challenges of integrating such systems into human social structures, including concerns about privacy, accountability, reliability, and the risks of excessive automation dependency. This scenario illustrates the challenges society faces in integrating Agentic AI into everyday life and in networks of interactions with other agents of all kinds.

13.2. Limitations

A primary benefit of Agentic AI lies in its ability to enhance efficiency and automate complex workflows, thereby liberating human resources for more strategic and creative endeavours. By utilising vast datasets and advanced algorithms, Agentic AI can identify innovative solutions to intricate problems, sometimes surpassing human capabilities. For instance, in software development, AI agents can manage workflows autonomously, adhering to established protocols and engaging in structured communication processes. This illustrates the potential of Agentic AI to streamline technical processes and improve operational efficiency. However, the implementation of social Artificial Agency faces significant technical constraints that shape its operational capabilities. Bandwidth limitations in distributed systems create bottlenecks that can impede performance, while computational resource allocation presents ongoing challenges in optimal system operation. The requirement for precise synchronisation and coordination across networks introduces additional complexity, and the reliance on digital infrastructure creates inherent vulnerabilities to technical failures and security breaches. Like Artificial Agency, Agentic AI struggles to cope with entirely new contexts outside its training parameters, having no semantic comprehension. This can lead to significant errors in unusual situations. Its ethical reasoning capabilities remain constrained by the training frameworks implemented during development, and its dependency on human-designed protocols and architectures creates inherent boundaries for autonomous evolution. A particularly challenging aspect lies in balancing individual optimisation with collective stability, especially given the compressed timeframes for evolution compared to biological systems, which have developed over millions of years.

Operational risks present another significant dimension of limitations. The potential for automated malicious activity poses serious security concerns, while the capacity for rapid misinformation propagation threatens information integrity. The growing reliance on these systems risks eroding human expertise and critical decision-making capabilities, potentially creating dangerous dependencies. The opacity of machine learning models can lead to unintended consequences that prove difficult to predict or control, and the potential reinforcement of biases present in training data continues to be a persistent challenge.

13.3. Implications

Social Artificial Agency challenges traditional theories of collective behaviour. It forces a reconsideration of anthropocentric views by demonstrating that sophisticated coordination and collective agency can emerge, or in this case, be designed, without consciousness, intelligence, biological imperatives, or evolutionary pressure. This insight suggests there may be universal principles underlying complex social systems, necessitating new theoretical frameworks that acknowledge artificial systems' distinct characteristics while differentiating them from biological ones.

As already indicated, deploying these systems creates unprecedented challenges in control, coordination, accountability, and governance, requiring innovative approaches to regulation and oversight. Traditional models of control, moral accountability/responsibility, and legal liability are inadequate when dealing with emergent behaviours arising from complex AI interactions. However, these challenges are balanced by significant opportunities for complex system management, offering novel solutions to previously intractable problems. The potential for hybrid forms of human-artificial social agency suggests new paradigms for collaboration between human and Artificial Agents, fundamentally altering our understanding of biological and human social systems. In particular, integration with human social structures will require careful management to optimise benefits while mitigating risks. As practical applications expand across economic and social domains, theoretical understanding becomes increasingly vital. The evolution of agentic AI challenges traditional agency concepts and opens new avenues for addressing complex societal challenges.

Beyond technological applications, social Artificial Agency provides insights into fundamental questions of intelligence, cooperation, and social organisation. Its dual role as both research subject and analytical tool makes it central to understanding collective behaviour and human-machine interaction paradigms. The future landscape of social Artificial Agency will likely be characterised by increasing integration with human systems, leading to hybrid forms of collective intelligence that combine the strengths of both artificial and human agency. This evolution suggests the potential emergence of entirely new forms of social organisation and problem-solving capabilities while raising important questions about autonomy, control, and the nature of intelligence itself. Understanding and managing this transformation represents one of the most significant challenges and opportunities in the ongoing development of artificial intelligence and its role in human society.

14. Conclusion

In this article, I argued for reconceptualising Artificial Intelligence (AI) as a novel form of agency rather than intelligence, proposing the Multiple Realisability of Agency (MRA) Thesis as the most comprehensive and coherent framework for understanding AI's nature, capabilities, and limitations.

By systematically comparing different forms of agency—natural, biological, animal social, artefactual, human individual, human social, and artificial—I identified distinctive characteristics and limitations of Artificial Agency while successfully avoiding anthropomorphic fallacies.

AI represents a distinct form of agency with programmable goals, data-driven adaptability, and distributed functionality. Unlike human agents, AI lacks consciousness, intentionality, and intelligence. However, its capacity for precision, scalability, and reproducibility enables it to excel in narrow, well-defined domains, making it a powerful complement to animal and human capabilities. I have also identified the emergence of social Artificial Agency (Agentic AI) as a novel collective form distinct from animal and human systems, opening up new possibilities for understanding and developing collective intelligence. These findings underscore the importance of integrating Artificial Agency with existing social structures through new theoretical frameworks and governance approaches.

Looking ahead, research will have to focus on developing agency-aware architectures that recognise AI's unique characteristics while creating robust governance frameworks for Artificial Agents and hybrid human-AI systems. The analysis suggests that advancing our understanding of emergent collective behaviours in artificial systems is crucial, as is exploring new paradigms for human-AI collaboration that leverage complementary capabilities. This development pathway requires careful attention to both theoretical understanding and practical implementation considerations. In this regard, I fully agree with (Ågerfalk 2020):

> "IS [Information Systems, my note] scholars should seize the opportunity to theorise agency in a digital context without relying on outdated (pre-digital) notions of institutions and agency that do not pay attention to the computational capabilities of information systems."

Indeed, there seems to be a need for a multi- and inter-disciplinary study of agency, not dissimilar but perhaps more successful than the field of cybernetics and complexity studies.

Examining the ethical landscape of Artificial Agency reveals significant opportunities and challenges. The potential for enhanced problem-solving capabilities through human-AI collaboration offers promising avenues for addressing complex societal challenges. Improved efficiency and effectiveness across various domains, coupled with the emergence of new forms of collective intelligence and social organisation, indicate positive transformative potential. However, these benefits must be weighed against risks of excessive automation dependency, challenges in maintaining meaningful human control, potential for unintended consequences in complex systems, and fundamental questions of accountability and responsibility. These issues have been recognised for decades in digital ethics. In the case of AI, governance frameworks must specifically address Artificial Agency rather than just applying traditional models that remain useful but potentially limited. Implementation should prioritise graduated autonomy systems that maintain appropriate human oversight, supported by clear accountability structures for hybrid human-AI systems. Ethical guidelines must recognise the unique characteristics of Artificial Agency while promoting beneficial integration. Continued research on human-AI interaction patterns and collective behaviour will prove essential for understanding and managing these evolving systems.

The future of Artificial Agency lies not in failing to replicate or surpass some kind of biological intelligence but in developing its unique agentic capabilities while ensuring alignment with human values and societal and environmental needs. This requires a careful balance between technological advancement and ethical considerations, supported by robust theoretical frameworks to understand AI development and practical governance approaches to regulate it. Success in this endeavour demands ongoing attention to the tremendous potential and significant responsibilities inherent in developing and deploying Artificial Agency systems. By carefully addressing these considerations, the development of an Artificial Agency can maximise its benefits, mitigate potential risks, and ensure alignment with human values, societal goals, and environmental sustainability.

Acknowledgements

Many thanks to Carlotta Buttaboni, Emmie Hine, and Jessica Morley for their comments on previous drafts.

References

Ågerfalk, Pär J. 2020. "Artificial intelligence as digital agency." *European Journal of Information Systems* 29 (1):1-8.

Albritton, Rogers. 1959. "On Wittgenstein's use of the term" criterion"." *The Journal of Philosophy* 56 (22):845-857.

Andrada, Gloria, Robert W Clowes, and Paul R Smart. 2023. "Varieties of transparency: Exploring agency within AI systems." *AI & society* 38 (4):1321-1331.

Bandura, Albert. 2006. "Toward a Psychology of Human Agency." *Perspectives on Psychological Science* 1 (2):164-180. doi: 10.1111/j.1745-6916.2006.00011.x.

Behdadi, Dorna, and Christian Munthe. 2020. "A normative approach to artificial moral agency." *Minds and Machines* 30 (2):195-218.

Bekoff, M. 2002. *Minding Animals: Awareness, Emotions, and Heart*: Oxford University Press, USA.

Chisholm, Roderick M. 1957. *Perceiving : a philosophical study*. Ithaca: Cornell University Press.

Dattathrani, Sai, and Rahul De'. 2023. "The Concept of Agency in the Era of Artificial Intelligence: Dimensions and Degrees." *Information Systems Frontiers* 25 (1):29-54.

Dennett, D. C. 1996. Kinds of minds : toward an understanding of consciousness. 1st ed, Science masters. New York, NY: Basic Books.

Dung, Leonard. 2024. "Understanding Artificial Agency." *The Philosophical Quarterly*.

Floridi, Luciano. 1997. "Scepticism and animal rationality: the fortune of chrysippus' dog in the history of western thought." *Archiv für Geschichte der Philosophie* 79 (1):27-57.

Floridi, Luciano. 2008. "The Method of Levels of Abstraction." *Minds and Machines* 18 (3):303-329.

Floridi, Luciano. 2013. "Distributed Morality in an Information Society." *Science and Engineering Ethics* 19 (3):727-743.

Floridi, Luciano. 2016. "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2083):20160112. doi: doi:10.1098/rsta.2016.0112.

Floridi, Luciano. 2023a. "AI as agency without intelligence: on ChatGPT, large language models, and other generative models." *Philosophy & Technology* 36 (1):15.

Floridi, Luciano. 2023b. The Ethics of Artificial Intelligence - Principles, Challenges, and Opportunities. Oxford: Oxford University Press.

Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines* 30 (4):681-694. doi: 10.1007/s11023-020-09548-1.

Floridi, Luciano, and Anna C. Nobre. 2024. "Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences." *Minds and Machines* 34 (1):5.

Floridi, Luciano, and Jeff W Sanders. 2004. "On the morality of artificial agents." *Minds and Machines* 14 (3):349-379.

Frankfurt, Harry G. 1971. "Freedom of the will and the concept of a person." *Journal of Philosophy* 68 (1):5-20.

Fritz, Alexis, Wiebke Brandt, Henner Gimpel, and Sarah Bayer. 2020. "Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI)." *De Ethica* 6 (1):3-22.

Gell-Mann, Murray. 1994. The quark and the jaguar: adventures in the simple and the complex. New York: W.H. Freeman.

Giddens, A. 1984. The Constitution of Society: Outline of the Theory of Structuration: Polity Press.

Haken, H. 1983. Synergetics: An Introduction: Nonequilibrium Phase Transitions and Self-organization in Physics, Chemistry, and Biology: Springer.

Himma, Kenneth Einar. 2009. "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11:19-29.

Kahneman, D. 2011. *Thinking, Fast and Slow*: Farrar, Straus and Giroux.

Kauffman, S.A. 1993. The Origins of Order: Self-Organization and Selection in Evolution: Oxford University Press.

Kauffman, S.A. 2002. *Investigations*: Oxford University Press.

Langley, Pat, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. "Explainable Agency for Intelligent Autonomous Systems." *Proceedings of the AAAI Conference on Artificial Intelligence* 31 (2):4762-4763.

Laukyte, Migle. 2017. "Artificial agents among us: Should we recognize them as agents proper?" *Ethics and Information Technology* 19:1-17.

List, Christian. 2021. "Group agency and artificial intelligence." *Philosophy & technology* 34 (4):1213-1242.

Luhmann, N. 1995. *Social Systems*: Stanford University Press.

March, J.G., and H.A. Simon. 1993. *Organizations*. 2nd ed: Wiley.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and information technology* 6:175-183.

Mayr, E. 1997. This Is Biology: The Science of the Living World: Harvard University Press.

Mintzberg, H. 2009. The Structuring of Organizations: A Synthesis of the Research: SSRN.

Norman, D. 2013. The Design of Everyday Things: Revised and Expanded Edition: Basic Books.

North, D.C. 1990. Institutions, Institutional Change and Economic Performance: Cambridge University Press.

Novelli, Claudio, Luciano Floridi, Giovanni Sartor, and Gunther Teubner. Forthcoming. "AI as Legal Persons: Past, Patterns, and Prospects." *Available* *at SSRN: http://dx.doi.org/10.2139/ssrn.5032265*.

Popa, Elena. 2021. "Human goals are constitutive of agency in artificial intelligence (AI)." *Philosophy & Technology* 34 (4):1731-1750.

Powell, W.W., and P.J. DiMaggio, eds. 2012. *The New Institutionalism in Organizational Analysis*: University of Chicago Press.

Prigogine, I., and I. Stengers. 1984. *Order Out of Chaos: Man's New Dialogue with Nature*: New Science Library.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four responsibility gaps with artificial intelligence: Why they matter and how to address them." *Philosophy & Technology* 34 (4):1057-1084.

Schreiber, Gerhard. 2024. "Reconsidering Agency in the Age of AI." *Filozofia* 79 (5).

Scott, W.R. 2013. Institutions and Organizations: Ideas, Interests, and Identities: SAGE Publications.

Sebastián, Miguel Ángel. 2021. "First-person representations and responsible agency in AI." *Synthese* 199 (3):7061-7079.

Shavit, Yonadav, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, and Alan Hickey. 2023. "Practices for governing agentic AI systems." *Research Paper, OpenAI, December*.

Simon, H.A. 1996. *The Sciences of the Artificial*. 3rd ed: MIT Press.

Swanepoel, Danielle. 2021. "Does Artificial Intelligence Have Agency?" In *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts*, edited by Robert W. Clowes, Klaus Gärtner and Inês Hipólito, 83-104. Cham: Springer International Publishing.

Swanepoel, Danielle, and Daniel Corks. 2024. "Artificial Intelligence and Agency: Tie-breaking in AI Decision-Making." *Science and Engineering Ethics* 30 (2):11.

Symons, John, and Syed Abumusab. 2024. "Social Agency for Artifacts: Chatbots and the Ethics of Artificial Intelligence." *Digital Society* 3 (1):2.

Taylor, C. 1989. *Sources of the Self*: Harvard University Press.

Tóth, Zsófia, Robert Caruana, Thorsten Gruber, and Claudia Loebbecke. 2022. "The dawn of the AI robots: towards a new framework of AI robot accountability." *Journal of Business Ethics* 178 (4):895-916.

van Lier, Maud. 2023. “Introducing a four-fold way to conceptualize artificial agency.” *Synthese* 201 (3):85.

Waal, F.B.M. 1982. Chimpanzee Politics: Power and Sex Among Apes: Harper & Row.

Weick, K.E. 1995. *Sensemaking in Organizations*: SAGE Publications.

Wellman, Carl. 1962. “Wittgenstein’s Conception of a Criterion.” *The Philosophical Review* 71 (4):433-447.

Wilson, E.O. 1975. *Sociobiology: The New Synthesis*: Belknap Press of Harvard University Press.

Wise, J. Macgregor. 1998. “Intelligent Agency.” *Cultural Studies* 12 (3):410-428.