DELIBERATIO
STUDIES IN CONTEMPORARY PHILOSOPHICAL CHALLENGES

https://deliberatio.uvt.ro/
2024, *4*, 55-69 SSN 2810 – 5826 ISSN – L 2810 – 5532 Print
ISSN 2810 – 5532 ISSN – L 2810 – 5532 Online

Artificial Intelligence and Moral Agency: Conceptual Clarifications for Philosophical Practice in AI Ethics

Robert Orgovan*

Abstract

The main objective of this paper is to explore the concepts of artificial agent and artificial moral agent, in a context in which the issue of ethics in the field of artificial intelligence (AI) is acquiring increasing relevance and urgency. The first part analyzes the concept of agent, together with the fundamental criteria proposed by Floridi and Sanders (2004) for defining an artificial moral agent, using the concept of "level of abstraction" (LoA) as a methodological instrument. The paper then presents the critique formulated by Johnson and Miller (2008), according to whom an artificial agent cannot possess moral status because of the absence of responsibility, the authors also raising objections concerning the inherent autonomy of such systems. In the final section, a series of the author's own theoretical considerations will be presented, accompanied by a structural proposal intended to bring greater rigor and clarity to the treatment of this issue.

Keywords: moral agent, artificial moral agent, responsibility, autonomy

1. What Is an Artificial Agent?

Among the first theorists to define the concept of artificial agent rigorously were Luciano Floridi and J. W. Sanders, in the article "On the Morality of Artificial Agents" (Floridi & Sanders, 2004). Before formulating a definition of the artificial agent, the authors introduce the concept of "level of abstraction," or LoA (Levels of Abstraction). In the text mentioned above, they explain that different levels of abstraction can exist with respect to the same entity. A level of abstraction consists of a collection of observable variables, each possessing a possible set of well-defined values (Floridi & Sanders, 2004, p. 354). A collection of levels of

* *Doctoral School of Philosophy, West University of Timișoara, Blvd. V. Pârvan 4, Timisoara (Romania). Email: robert.orgovan@e-uvt.ro. ORCID: 0000-0009-0003-8086-0704.*

abstraction forms an interface, which is then used to analyze a system from multiple perspectives. The analysis of a system through this interface leads to the generation of a model.

To illustrate the mechanism of levels of abstraction, the authors propose a hypothetical scenario: a conversation among three interlocutors, x, y, and z. Although the exact subject of the discussion is not known, the participants' profiles are specified: x is a collector and potential buyer; y has technical skills and spends his free time maintaining mechanical equipment; and z is an economist. During the dialogue, the following observations are made: x notes the presence of an anti-theft device and the fact that the object has had only one owner; y observes that the engine is not the original one, that the leather components show wear, and that the bodywork has recently been repainted; while z emphasizes the high cost of spare parts, while also mentioning that the object retains a stable market value (Floridi & Sanders, 2004, p. 354). This example highlights the importance of the perspective from which each observer evaluates the object. Regardless of the exact nature of the entity under discussion (which appears to be an automobile or a motorcycle), each participant analyzes the object from the standpoint of his or her own field of interest and therefore operates with different levels of abstraction. The intersection and cumulative combination of these three perspectives (the three LoAs) constitute the interface. It is also important to emphasize that one LoA can include another; in this case, the level of abstraction that encompasses the other is considered to be at a lower level, that is, a more concrete one. Continuing the example, if a new participant (Ana) were to enter the discussion, and if her LoA included the perspectives of x and y, her level of abstraction would be considered lower. This is because x's LoA is more abstract in comparison with Ana's, since Ana operates at a more generalized level, not being interested in, and not knowing, the specific abstract details analyzed by x and y (Floridi & Sanders, 2004, p. 355).

The concept of agent is defined as a system embedded in an environment, capable of initiating a transformation, producing an effect, or exerting an influence on that environment over time (Floridi & Sanders, 2004, p. 357). This point underscores the need to make use of the concept of level of abstraction. According to this preliminary definition, even an extreme natural phenomenon (for example, a hurricane or an earthquake) could be classified as an agent. This perspective corresponds to a primary level of abstraction, which operates with a limited set of properties. For this reason, Floridi and Sanders argue that the level of abstraction must be lowered, the level appropriate to the analysis being the one that presupposes the simultaneous presence of three fundamental properties: interactivity, autonomy, and adaptability. The selection of these criteria is not arbitrary; the level of abstraction in question must reflect those features that characterize interaction with the external environment and with other agents. At this level of analysis, the human being is conceptualized not merely as a simple agent, but as a moral one.

Interactivity presupposes the existence of reciprocal action between the agent and its environment. This process generally takes place through input and output data relations at the level of the system, or through an action in which the agent, the environment, or another agent within that environment are involved simultaneously. An illustrative example in this respect is the interaction between a person and his or her dog during play. Autonomy designates the agent's capacity to modify its internal state without receiving a direct command and without representing a strictly reactive response to an interaction. This property gives the agent a certain degree of independence and complexity in relation to its environment. The example of the dog is applicable in this context as well, since the animal demonstrates the autonomy required to carry out various activities without depending on a specific input for each individual action. Adaptability indicates the agent's ability to modify its set of operational rules in order to respond effectively to novel situations. This implies that the agent is capable (relative to a given LoA) of evaluating its own mode of operation and adjusting it when necessary, through a process of learning based on experience. The use of the example of the dog is not accidental, since it satisfies all three criteria simultaneously. Depending on the context in which they find themselves, dogs demonstrate the capacity to adapt by modifying their behavioral states. Therefore, on the basis of this analytical framework, it can be stated with certainty that a dog constitutes a valid agent.

Up to this point, the concept of agent has been defined by appealing to levels of abstraction; nevertheless, the strict application of this conceptual framework to artificial agents encounters certain theoretical limitations. To illustrate this problem, Floridi and Sanders turn to the example of the MENACE system (Matchbox Educable Noughts and Crosses Engine). MENACE is a mechanism made up of a multitude of matchboxes, each box corresponding to a possible state of the board in the game of noughts and crosses. Each available space on the board is associated with a specific color, and in each box there are initially placed three or four beads of each corresponding color. When a human player makes a move, the box corresponding to the new configuration of the game is identified; the box is shaken, and a bead is randomly extracted, thereby indicating the move selected by the system. Its functioning is based on a conditioning mechanism: if the system is defeated, the beads extracted and used in that match are removed, thereby reducing the probability that the same choice will be repeated; in the event of a victory, one bead of the color used is added to each activated box. The MENACE system functions as a mechanism capable of learning from its own errors. It requires a considerable number of iterations (matches played against an opponent) in order to reach a performance threshold at which defeat becomes impossible. Its architecture aims to simulate the human reasoning required to achieve victory. From an operational point of view, MENACE functions under two rules: it must block the opponent from completing a horizontal, vertical, or diagonal line, while simultaneously aiming to complete one of these lines itself, under the constraint

that a square cannot contain two different symbols. Given these operating parameters, the system must assimilate winning strategic patterns, a process that is realized strictly through repetition. This example is used to examine the different levels of abstraction (LoAs) applicable to the MENACE system (Floridi & Sanders, 2004, pp. 359–362). The first level of abstraction analyzed is that of the individual match. At this level, only the state of each round within the match and its outcome can be observed. In this restricted context, the system demonstrates interactivity (by providing an output for every input received) and autonomy (because, for each position, the choice is determined by probabilities influenced by the history of matches played and won, generating different combinations). However, the system does not manifest adaptability. At the level of a single match, the decision-making process behind the moves remains opaque, since there is no observable basis from which one could infer that the system learns from experience. Therefore, strictly in relation to this LoA, MENACE does not meet the criteria for being considered an agent. The second level of abstraction is that of the tournament. Within this analytical framework, a significant number of successive matches are observed, together with their corresponding results. As at the previous level, MENACE satisfies the criteria of interactivity and autonomy, but this time the criterion of adaptability is also added. The extended series of games provides an observable database sufficient to support the conclusion that the system has learned from experience and adjusted its behavior from one match to another. Consequently, at the second level of abstraction, MENACE can be classified as an artificial agent. The third level of abstraction is that of the system. In this analytical framework, one observes not only the sequences of matches and the results obtained by MENACE, but also directly examines the source code and the fundamental mechanisms of operation. As at the previous levels, the system retains the attributes of interactivity and autonomy, but it loses the status of adaptability. This loss occurs because, once access is obtained to the source code or to the instruction manual, it becomes evident that what at a higher level appeared to be adaptability is, in essence, a simple deterministic updating of the program. If a transition rule is exclusively a consequence derived from the internal state of the program, the process can no longer be defined as adaptability. Despite its probabilistic character, the transition rule functions strictly as a predetermined reaction to a specific input. Therefore, at this level of abstraction (LoA), MENACE no longer satisfies the conditions for being considered an agent.

The example proposed by Floridi and Sanders introduces a significant degree of conceptual ambiguity, generating an apparent paradox: MENACE both is and is not an artificial agent, depending on the perspective of analysis. First, the rigorous application of this methodology based on levels of abstraction could lead to the conclusion that, at an extremely low level of abstraction (reduced to physicochemical processes), not even the human being could still be considered an agent. One may invoke, in this respect, the theoretical scenario in which human

emotional states, such as anger or fear, are strictly the result of deterministic neurochemical reactions, without which the attainment of those states would not be possible. Second, following this logic, regardless of the degree of complexity of a program or the sophistication of its cognitive architecture designed to fulfill tasks requiring an equivalent of human reason, analysis at the level of the source code would invariably invalidate its status as an agent. In order to overcome these theoretical problems, it is necessary to introduce the perspective formulated by Johnson and Miller (Johnson & Miller, 2008), who subjected the theoretical foundations proposed by Floridi and Sanders to critical analysis.

2. What Is a Moral Agent? Is an Artificial Moral Agent Possible?

2.1. Floridi's Perspective on Artificial Moral Agents

Before defining what a moral agent is, the concept of moral action must be defined. According to Floridi (2013), an action is considered moral if and only if it has the capacity to generate moral good or moral evil, more precisely if it "decreases or increases the degree of metaphysical entropy in the infosphere" (Floridi, 2013, p. 147). The infosphere is conceptualized as an analogue of the biosphere, representing a metaphysical dimension of information, knowledge, and communication, populated by informational entities (Floridi, 2013, p. 28).

To illustrate this paradigm, Floridi proposes a thought experiment involving two entities, H and W. In this scenario, H initiates a lethal action against a person, while W acts in order to heal that person. Both entities interact with elements of their environment and demonstrate autonomy, possessing the theoretical capacity to opt for alternative decisions and modifying their internal state in the course of the action. At the same time, both H and W display adaptability, acting independently of predetermined instructions. It can be postulated that both entities have adjusted their behavior according to environmental variables, thereby optimizing their individual intervention. The problem that follows from this is the following: on the basis solely of this behavioral description, how can the human agent be distinguished from the artificial one?

This problem can be reevaluated through the lens of the Turing test, extended here in the form of a "moral Turing test." In this context, a series of general objections arise, similar to those originally formulated by Turing in his work. Floridi analyzes in extenso four of these counterarguments, focused on the boundaries between the human being and moral artificial intelligence: the teleological objection, the intentionality objection, the objection concerning freedom of action, and the responsibility objection (Floridi, 2013, pp. 148–152).

The teleological objection postulates that an artificial agent lacks its own finality, not being intrinsically oriented toward a goal. This argument can easily be rejected, since artificial intelligence systems can be designed and optimized to

manifest behaviors strictly oriented toward the achievement of specific objectives. Therefore, the capacity to formulate and attain a goal does not constitute a viable criterion of demarcation between the human being and artificial intelligence. The intentionality objection maintains that an artificial system cannot possess intentional states. According to this perspective, for an action to be considered moral, the agent would have to relate to its acts reflexively, implying the existence of meaning, a desire to act, or self-awareness.

Although the objection is valid from an ontological perspective, it loses its relevance in the strict evaluation of a moral agent. All agents in an interactive system participate in what is called the "moral game"; nevertheless, awareness of this participation or the deliberate intention to participate is not a determining factor. The crucial aspect is the nature of the action itself and its potential to generate objective moral consequences.

Thus, intentionality does not represent a sufficiently strong discriminatory criterion for separating human moral agency from artificial moral agency in this analytical register. The objection concerning the problem of freedom maintains that an artificial intelligence does not possess the free will specific to the human being.

This criticism can be rejected by an argument similar to the preceding one. Any objection that targets exclusively the internal states specific to human psychology is irrelevant in the context of the present analysis. Although these states mark a real difference between the human being and artificial systems, at the level of moral agency they do not constitute a functionally discernible distinction. The "freedom" of artificial intelligence systems derives from their non-deterministic nature. A similar counterargument was also formulated by Turing in addressing these types of objections.

The final objection, focused on responsibility, postulates the incapacity of an artificial agent to be held accountable for its actions. This is considered the only strong objection, since there would be no logical basis for praising, rewarding, blaming, or punishing an artificial entity. In this sense, Floridi formulates two observations concerning the concept of responsibility. The first observation highlights the human reluctance to attribute blame to artificial agents, a reluctance grounded in the absence of a psychological component in the machine. Nevertheless, after negative consequences occur, these agents can be classified as sources of evil; in this case, the act of redesigning or reprogramming the system represents the functional equivalent of "punishment," being an action similar to the process of sanctioning or rehabilitating a human individual who has committed a reprehensible act. The second observation derives from the premise that the artificial agent is exempted from responsibility by human evaluators. In a first sense, one may maintain that an artificial system does not bear responsibility for an act if it operates with an incomplete set of data about its environment or if it is unable to identify an alternative solution to a given problem. In such circumstances, regarding the artificial agent as morally responsible is an unfair undertaking.

Consequently, the level of abstraction (LoA) should integrate an additional criterion—alongside the preexisting ones—for an agent to acquire moral status: responsibility. Ultimately, it is precisely the absence of this responsibility that constitutes the essential line of demarcation between the human agent and the artificial one.

Nevertheless, the objection based on the concept of responsibility also presents certain theoretical vulnerabilities. Although it is traditionally postulated that an agent must bear responsibility for its actions, empirical reality offers significant counterexamples. To illustrate this asymmetry, Floridi invokes two telling examples: young children and working dogs (search-and-rescue dogs). In the case of a child, although the child may constitute the causal source of harm, legal and moral responsibility falls to the parents. By analogy, in the case of an artificial agent, the role of the "parents" could be attributed to the programmers or designers of the system. The situation acquires an entirely different complexity, however, in the case of search dogs. These animals are trained to locate missing persons, thereby participating directly in what can be defined as a "moral game." Although human beneficiaries frequently develop attachment and gratitude toward their canine rescuers, for the latter the entire operation represents exclusively the execution of a training pattern assimilated to a game. Although these dogs cannot be considered responsible for their actions (which have positive moral effects), human observers tend to attribute to them the quality of moral agents, based primarily on their central role in carrying out the rescuing action. The validity of this example demonstrates that responsibility functions, in certain contexts, as a "pseudo-criterion" in defining the moral agent. A significant distinction proposed in this analysis must nevertheless be emphasized: unlike other types of agents, the search dog is structurally limited in its capacity to participate in scenarios with immoral implications. The evaluation therefore acquires a biased character: the animal is invested with moral status precisely because the probability that it will perform a harmful action during the mission is negligible (for example, it is unlikely that a search dog will attack the person it locates). This conceptual asymmetry reveals a contradiction in the way morality and responsibility are attributed. There is a tendency to confer moral status and responsibility on a non-human agent when it produces positive effects (out of subjective feelings of gratitude), while denying it the quality of a responsible moral agent when it has the potential to generate immoral consequences (for example, because of algorithmic errors). The prejudice according to which only the human being is capable of moral action leads to contradictory approaches. Either the moral act of an artificial agent is reduced to mere structural determinism (the programmed code), or, under the sway of emotions, these mechanisms are ignored and the attribute of "morality" is irrationally assigned. Thus, in such cases, ethical discourse frequently and unjustifiably oscillates between two methodological extremes.

After formulating responses to the four objections, Floridi introduces the concept of a threshold function applicable to a given level of abstraction (LoA). An agent evaluated at this LoA is considered morally beneficial if and only if, in relation to a predetermined value called "tolerance," it maintains a functional relationship with the observable elements of the environment, such that the value of the threshold function never exceeds the admitted tolerance limit. In the case of an artificial agent, the threshold function is determined algorithmically through a formula; however, tolerance is identified and calibrated by human agents (although, in principle, it can be determined a priori), the latter exercising the fundamental ethical judgments.

This paradigm makes possible the ethical analysis of various types of artificial agents, such as a thermostat, an algorithm (bot) for filtering electronic correspondence, or the MENACE system. Both the thermostat and the filtering bot possess an inherent moral charge, since both influence, to one degree or another, human well-being. In the case of the thermostat, its action is considered morally correct if the operational result maintains a level of comfort within the tolerance limits agreed upon by the user. In this context, the threshold can be conceptualized as the gap between the actual state of the environment and the optimal state desired by the human being. Analogously, for the email-filtering bot, the threshold could be defined as the percentage of messages classified incorrectly.

With regard to the MENACE system, evaluating its actions in terms of morality requires the introduction of additional criteria. Inherently, the primary function of this program is to play games of noughts and crosses, an axiologically neutral activity that does not make it eligible for beneficial or harmful acts in the absence of a specific evaluative framework. Nevertheless, parameters can be associated with it that confer moral valences on its behavior, one such criterion being the calibration of difficulty in relation to the experience of play. For example, MENACE would manifest unethical behavior if it allowed the human opponent to win easily; the consequences of this conduct would consist in depriving the player of the satisfaction of a deserved victory and in creating an illusion of the player's own competence, thereby transforming the software into a deceptive agent. One way of modeling a threshold function in this case would be to relate it to the number of games required for the user to defeat the system. Once this threshold is reached, the software should increase its degree of difficulty, evolving toward invincibility. Thus, if the algorithm establishes that 25 matches are required to defeat MENACE, at match number 26 the system will adjust its difficulty upward.

In conclusion, Floridi reanalyzes the concept of the responsibility of the artificial agent and the way in which it should be managed. Frequently, in the case of errors produced by an artificial intelligence, blame tends to be attributed to the persons who created the system (the team of programmers, system engineers, testing engineers, and so on). At the same time, eligibility for blame is not limited to technical personnel, but can extend equally to administrative and executive

branches. For this reason, Floridi (2013, pp. 156–157) proposes a mechanism of "censorship" for immoral artificial agents, structured in three stages. The first stage consists in monitoring and modifying the agent (maintenance and reconfiguration operations); the second stage involves isolation, namely the removal of the agent from the disconnected component of the infosphere; the final stage presupposes the definitive elimination of the agent from the infosphere. This approach suggests that the systemic treatment of agents—using the principles of an antivirus program—is a far more beneficial and efficient solution than the process of searching for the guilty "creator."

2.2. Johnson and Miller's Critique of Floridi and Sanders's Perspective

Four years after the publication of the article by Floridi and Sanders, Johnson and Miller formulated the first rigorous rejection of the concept of artificial moral agents (Johnson & Miller, 2008). The authors begin their argument with an analysis of the term agent (especially the artificial agent), a move followed by a critical evaluation of its moral dimension. In their study, they classify the theorists who investigate the status of the artificial agent into two broad categories: the group of "Computational Modelers" and the "Computers-in-Society Group." The first group approaches computation either as a model for representing reality or as a tool for scientific exploration, a category that also includes the approach proposed by Floridi and Sanders (Johnson & Miller, 2008, p. 126). Supporters of this perspective regard computational modeling as the optimal solution for understanding reality, relying on the premise that reason leads directly to truth. Thus, this group uses computation as an epistemological foundation for a body of knowledge capable of generating new perspectives across a vast range of fields. The Computational Modelers designate information systems as agents, starting from the hypothesis that, from a behavioral point of view, they can function similarly to human beings. Moreover, they postulate that computer systems can be used not only to model, but also actively to generate behavior (Johnson & Miller, 2008, p. 126). An illustrative example in this respect is the evolution of stock-exchange trading. Sequential operations (such as checking the price of a share or completing documentation), previously carried out by a multitude of human agents, are now taken over by a computer system. The latter is configured by a single human agent to buy or sell shares autonomously, exclusively on the basis of predetermined parameters.

The second group investigates the role of technology in the lives of individuals, with particular interest in the moral dimension of this interaction. The central objective of this group is to sensitize the public to the disruptive force of technology and to its ethical implications, an undertaking carried out by revealing the inherent moral character of information systems and the values embedded in

their design process. Johnson and Miller place themselves within this category (Johnson & Miller, 2008, p. 126). Although the group recognizes the behavioral independence of information systems as a source of ethical concern, it rejects their classification as autonomous moral agents. The main argument is pragmatic in nature: attributing moral autonomy to information systems would risk freeing their designers from legal and moral responsibility for the impact of those systems on human well-being. While the first category is grounded in formal logic and the philosophy of mind, the second group derives its premises from social philosophy and applied ethics (Johnson & Miller, 2008, pp. 126–127).

The core of the debate between the two perspectives concerns the way in which computer systems that manifest a degree of independence are constructed. A theoretical victory for the first perspective would confer a distinct moral status on artificial agents, which could hypothetically lead to the imposition of prohibitions on deactivating them. By contrast, the perspective of the Computers-in-Society Group maintains that although artificial agents are relevant elements of the moral space, they must remain under permanent human control. From this perspective, systems are understood exclusively as instruments designed to serve human purposes, whose operation is impossible in the absence of constant human supervision. After setting out these currents, Johnson and Miller formulate a substantial critique of the method of levels of abstraction (LoA) proposed by Floridi and Sanders. The fundamental critique (Johnson & Miller, 2008, pp. 126–127) concerns the high degree of methodological relativity, manifested in the difficulty of determining objectively when an agent can be considered independent or moral. From a logical point of view, it is problematic that an agent can possess property p at one level of abstraction and, simultaneously, property ~p at another level, a fact that generates a contradictory state. In order for a computer system to be validated as an artificial moral agent, the authors maintain that a rigorous delimitation of the concept of "moral" would be necessary within the specific LoA in which the system is identified as autonomous.

Finally, a pertinent critique concerns the way in which the discourse on "artificial moral agents" may in fact conceal a series of influence groups and economic or political interests. Considerable resources—financial, temporal, and human—are directed toward the development of technological infrastructures, weapons systems, and monitoring mechanisms whose ultimate purpose is to serve specific human interests (such as optimizing security, increasing the efficiency of communication, or increasing productivity).

Thus, the concept of artificial moral agent inherently conceals a set of anthropocentric intentions. In this context, perspectives such as that of Bill Joy (Johnson & Miller, 2008, p. 131) warn against the possibility that artificial agents might reach a level of sophistication that would allow them to take control over humanity. Such a hypothesis is sustainable only by accepting the idea of

technological determinism, a paradigm in which the evolution of technology follows an implacable trajectory removed from human control.

By contrast, the view of the Computers-in-Society Group maintains that technological evolution is continuously governed by human decisions; therefore, critical monitoring of design and construction processes is imperative, while a passive attitude toward technological progress is considered inadvisable. These arguments constitute the core of Johnson and Miller's critique of the concept of artificial moral agency. Although their objection regarding the theoretical validity of the "levels of abstraction" (LoA) proposed by Floridi and Sanders is well founded, the general criteria for defining an agent can still be considered heuristically valid. The major difficulty nevertheless persists in defining the ethical dimension of these agents, an aspect in which Floridi and Sanders's theory does not succeed in offering a fully convincing explanatory framework. Although the specialized literature has subsequently registered multiple similar objections, Johnson and Miller's study remains the cornerstone of this critical paradigm.

A convergent approach is offered by Patrick Hew (2014), who, in agreement with the representatives of the Computers-in-Society Group, reiterates the idea that responsibility remains a strictly human prerogative. He argues that an artificial agent could be considered worthy of moral appraisal only in the scenario in which its behavior, as well as the generation of its operating rules, were completely independent of external human programmers. In the literature devoted to the philosophy of moral action, analysis frequently focuses on systems based on compliance with a set of predetermined rules. Hew (2014, p. 198) uses the analogy of the mousetrap to illustrate the conceptual limits of this perspective: "Continuously monitor pressure on the trigger. If the pressure exceeds the threshold, then activate the trap." Such a mechanism is completely devoid of decisional responsibility and is therefore impossible to qualify as a moral agent.

Through this reasoning, Hew demonstrates that no artificial agent, regardless of its degree of sophistication, can be the bearer of genuine moral responsibility, any such attribution constituting, in essence, a "false" responsibility that invalidates the very idea of an artificial moral agent. The author also extends his critique to cybernetic systems characterized by self-regulation, self-replication, and self-modification—systems that partially simulate independence through the use of learning algorithms. Despite this apparent operational autonomy, the fundamental architecture of the system (the initial source code) remains a human construct, functioning on the basis of primordial rules implemented for precise purposes by its creators.

3. The Author's Own Perspective on the Problem and Conclusions

As for defining the concept of an agent, the formulation proposed by Floridi and Sanders remains rigorous and widely applicable in contemporary research. The

three fundamental criteria—interactivity, autonomy, and adaptability—accurately describe the natural relationships between various biological organisms, which we intuitively recognize as agents. The major difficulty of their theory lies, however, in the concept of level of abstraction (LoA), which proves to be an excessively relative variable and difficult to delineate, particularly in the evaluation of the autonomy criterion. In the absence of a clear demarcation line between levels of abstraction, we arrive at the paradoxical situation in which certain prototypes can be classified simultaneously as both agents and non-agents. Compared to human or biological agents, artificial prototypes require a distinct set of criteria for evaluating autonomy, since the decisions of organic agents are also influenced by inherent impulses or instincts. A prototype achieves the level of autonomy necessary to be validated as an agent when it demonstrates the capacity for memory and learning (including from its own errors), and its mode of operation transcends strict determinism. More specifically, the results are not conditioned exclusively by parameters or data provided a priori, but are generated dynamically, either through algorithmic reasoning or through the independent extraction of data from sources. More specifically, the results are not determined solely by parameters or data provided in advance, but are generated dynamically, either through algorithmic reasoning or by independently extracting data from external sources (databases).

An artificial agent that does not satisfy the conditions of strong artificial intelligence will never possess the level of autonomy specific to the human being. Nevertheless, if such a system is capable of generating one or more relevant outputs after processing complex sets of data, it can be said to possess a certain degree of operational autonomy. In this context, human cognitive processes are replaced by algorithmic architectures. Taking as a model the taxonomies that classify the degrees of independence of autonomous vehicles, it becomes appropriate to create a similar universal system for artificial agents. This paper proposes an evaluative system structured on six levels of autonomy for prototypes, arguing that only systems that fall between levels 3 and 5 can legitimately be considered artificial agents.

Level zero of autonomy comprises prototypes whose software architectures are entirely programmed and predetermined by the human operator. These systems execute exclusively the instructions received, operating on the basis of a rigid set of parameters. Clear examples in this respect are electric elevators or elementary entertainment systems (such as automated toys capable only of reproducing or mechanically paraphrasing an audio signal). Consequently, a level-zero prototype does not acquire the status of agent, being, in essence, a simple instrument intended for practical utility or entertainment. Level one of autonomy includes prototypes that execute predefined sets of instructions but query and use external databases in order to complete tasks. Systems situated at this level do not possess a decision-making capacity proper; their primary function consists in processing and

displaying sets of data in real time. A representative example is a computer program designed to monitor air traffic, capable of reproducing the coordinates and flight schedules of aircraft in real time. Although ontologically close to level zero, this type of system manages a considerably larger volume of information, which it sorts and structures according to specific queries. Despite this increased informational complexity, level-one prototypes do not qualify as agents.

At level two of autonomy are found prototypes which, although they function on the basis of general instructions, use sensory inputs to operationalize the task, involving an analytical process of solving spatial or contextual problems. Appropriate examples for this category are service robots (waiter-type robots) or semi-autonomous photographic drones. In the case of the service robot, although the basic command is strictly formulated (for example, transporting object X to destination Y) and the map of the environment is predefined, the system must actively manage the avoidance of unforeseen obstacles along the route. On the basis of the preset topography, the system calculates the optimal trajectory and is capable of reconfiguring its route in real time when a physical impediment is detected. In addition, it may possess facial-recognition modules for the correct identification of the user, similar technologies being applicable in the case of drones as well.

At this stage, a significant qualitative leap in autonomy is registered; nevertheless, dependence on a predetermined map and the absence of a generalized learning algorithm for receiving commands limit the system to the status of a primitive agent, whose actual learning capacity is strictly limited to optimizing navigation routes. This level marks the transition point from a simple operational prototype to an agent proper, since the degree of autonomy attained eliminates the need to program manually every step in the movement process.

Level three of autonomy includes prototypes capable of taking up a task and executing it with complete operational independence from the human operator. A paradigmatic example is the autonomous vacuum cleaner, which, once activated, initiates a series of algorithmic routines in order to fulfill its singular objective (cleaning the space). Equipped with a network of detection sensors, the system adapts to the surrounding environment and autonomously generates the spatial mapping required to carry out the task. Moreover, it assimilates the routes and obstacles encountered through a process of reinforcement learning (trial and error). In future iterations, the system memorizes the coordinates of objects; in the event of a change in the topography of the space, the algorithm dynamically recalculates the optimal trajectory. At this level, the prototype unquestionably acquires the status of an artificial agent, since it simultaneously satisfies both the criteria of autonomy (the capacity for learning, memorizing, and processing data) and the fundamental properties of an agent (autonomy, adaptability, and interactivity).

Level four of autonomy brings together prototypes that possess the characteristics of level three, the demarcating factor consisting exclusively in the

weight and ethical importance of the decisions adopted. This category may include "moral agents," that is, those systems whose decisions generate consequences with considerably more severe axiological and physical impact. This level plays a taxonomic separating role, justified by the impossibility of establishing an ethical equivalence between a domestic utility system (such as the autonomous vacuum cleaner) and, for example, an autonomous vehicle.

Although both systems operate with a similar degree of algorithmic independence, level-four systems require human supervision until their efficiency in critical scenarios (edge cases) is statistically validated. This category includes autonomous vehicles, security systems based on artificial intelligence (artificial guards), and algorithms for penalizing user behavior in virtual environments. Level five, the highest stage of this hierarchy, corresponds to the concept of "strong artificial intelligence" (Strong AI), postulated by Searle (1980). Such an agent would manifest a cognitive independence equivalent to that of the human being, representing the theoretical ideal and the terminus of technological development in the field.

Following this classification, it is necessary to clarify the concept of responsibility and its necessity in validating the moral status of a prototype. The present analysis maintains that responsibility, in its traditional (human) sense, is not an obligatory criterion. Attributing responsibility to an artificial agent represents a category error, justified by the impossibility of applying punitive or morally corrective measures to a computer system. Final responsibility therefore belongs to the human factor (the creators or designers). Nevertheless, situations arise in which not even the human network can bear absolute responsibility. For example, the unexpected malfunction of a sensor while an autonomous vehicle is running does not necessarily imply fault on the part of the design engineers, since the causes of the malfunction may stem from unforeseeable environmental factors. Therefore, the optimal technical solution consists in integrating fail-safe mechanisms (self-diagnosis); upon detecting even the smallest anomaly, the system must be programmed to interrupt its activity under conditions of maximum safety (e.g., by automatically initiating a parking maneuver). In addition, the hardware of systems with major risk potential must meet the highest standards of redundancy and technological resilience.

In the situation in which the agent is delegated to make critical decisions (involving life and death), algorithmic errors transfer blame directly to the human network. It can be argued that, except for defects in the physical infrastructure or direct errors in writing the code, legal and moral responsibility belongs primarily to the executive and administrative decision-makers who approved the implementation and commercialization of the system in question. Although attributing direct responsibility to the artificial agent proves to be a useless undertaking, this limitation does not annul the system's capacity to generate decisions of a moral character. Therefore, morality can be attributed to artificial

agents strictly in the form of an operational functionality: they perform actions with moral implications (beneficial or harmful) without, however, possessing an intrinsic morality ("in the strong sense") grounded in consciousness or free will.

References

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.
https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing Test. *Journal of Information, Communication and Ethics in Society, 13*(2), 98–109.
https://www.researchgate.net/publication/276463060_Issues_in_robot_ethics_seen_through_the_lens_of_a_moral_Turing_test

Greco, G. M., & Floridi, L. (2004). The tragedy of the digital commons. *Ethics and Information Technology, 6*(2), 73–81.
https://link.springer.com/article/10.1007/s10676-004-2895-2

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3), 197–206.
https://link.springer.com/article/10.1007/s10676-014-9345-6

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2), 123–133.
https://link.springer.com/article/10.1007/s10676-008-9174-6

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.
https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://academic.oup.com/mind/article/LIX/236/433/986238