*Φιλοσοφια*
**Volume 24, 1: 2023**

# ACTION AND AGENCY IN ARTIFICIAL INTELLIGENCE: A PHILOSOPHICAL CRITIQUE

**Justin Nnaemeka Onyeukaziri**
***Fu Jen Catholic University, Taiwan***

*The objective of this work is to explore the notion of "action" and "agency" in artificial intelligence (AI). It employs a metaphysical notion of action and agency as an epistemological tool in the critique of the notion of "action" and "agency" in artificial intelligence. Hence, both a metaphysical and cognitive analysis is employed in the investigation of the quiddity and nature of action and agency per se, and how they are, by extension, employed in the language and science of artificial intelligence. The advent of the science of artificial intelligence and cognitive science, and the technological applications of artificial intelligence in the production of agents such as driverless cars and expert systems, have raised the question of moral, ethical and/or legal responsibility in AI agents. This has re-emphasized the importance of the philosophical discourse on the notions of action and agency, which in contemporary intellectual discourse are now perceived to be phenomena within the epistemic competence of the natural sciences. This paper argues that AI systems do not and cannot possess free agency and autonomy, thus, cannot be morally and ethically responsible. Hence, it recommends a socio-political response to the question of responsibility in AI. It is then the duty of individual nations, or the global community to define and enact policies on who shoulders the responsibility of actions executed by AIs.*

*Keywords: Action, agency, Artificial Intelligence, person, mind, consciousness*.

## INTRODUCTION

With the advancement in the development of Artificial Intelligence (AI), both as the science of intelligence, and as technology that simulates human intelligence in decision-making and problem-solving, the questions on the moral, ethical,[1] and legal responsibilities of AI actions are becoming more frequent and prominent. These questions are definitely outside the epistemic framework and authority of the science

of AI. Philosophy is rather the proper science, whose nature is to reflect and elucidate the epistemic problematics around the moral, ethical, and even legal questions around the science and development of AI. The questions of moral, ethical, and legal responsibilities in AI presuppose that AIs are intelligent and can execute actions that could be blame-worthy (or praise-worthy). If thus, then they are intelligent systems, otherwise known as agents. If they are agents, it presupposes they have or could have autonomy. Hence, the need to expound the notion of "action" and "agency" in AI. This need calls for a philosophical critique of the notion of "action" and "agency" in AI (intelligent computers/systems) which necessarily entails the question of autonomy.

Therefore, this work aims to explore the notion of action and agency in artificial intelligence. It employs a metaphysical notion of action and agency as an epistemological tool in the critique of the notion of "action" and "agency" in AI. The metaphysical investigation will be based on the Aristotelian-Thomistic tradition. Hence, both a metaphysical and cognitive analysis shall be employed in the investigation of the quiddity and nature of action and agency per se, and how it is, by extension, employed in the language and science of artificial intelligence. The advent of the science of artificial intelligence and cognitive science, and the technological applications of artificial intelligence in the production of "agents" such as driverless cars, expert systems, AI-designed robots, and drones, have re-emphasized the importance of the notions of "action" and "agency" in philosophical discourse. The advent of AI and Cognitive Science research in contemporary sciences that have become more materialistic or naturalistic make the notions of "action" and "agency" to be perceived as phenomena within the epistemic competence of the natural sciences. These notions are employed in the language of computer science, AI, and cognitive science, in a manner similar to their usage with respect to the human person. However, when an intelligent system is said to have carried out an action or that it is an agent is the notion of "action" and "agent" or "agency" employed to have the same conceptual content as when they are employed in respect to the human person? And do they have the same moral, ethical and/or legal implications they contain as when they are employed with respect to the human person? For a comprehensive response to these problematics, it is, therefore, necessary to philosophically reflect on the metaphysics of action and agency and how these notions relate to the philosophy of mind and human subjectivity and autonomy. This is the overall objective of this work. This work is executed thus: Introduction, The Metaphysics of Action and Agency, Action and Agency in Artificial Intelligence, Philosophical Critique of Action and Agency in AI, and Conclusion.

## THE METAPHYSICS OF ACTION AND AGENCY

Metaphysics of action and agency aims to investigate the quiddity, nature[2] and end of action and agency. Action and agency seem to deal with the same faculty in the human person. They both involve the faculty of the will. However, they do not exclusively involve the will; they also involve the intellect since action and agency

involve knowledge or the act of intellection. In the Aristotelian-scholastic tradition, the will and the intellect are powers of the soul.[3] Aristotle is probably the first philosopher in the West to give a philosophical reflection and analysis on the notion of action and agency. According to Aristotle's logic, physics, and metaphysics, action (or active, of which contrary is passive) is one of the categories of being (substrate). According to Aristotle (Metaphysics 1021a15): "The active and the passive imply an active and a passive capacity and the actualization of the capacities." The active is that which acts, while the passive is that which is acted on. The active and the passive could also be understood in the sense of "the agent" and "the patient." The agent is that which moves (acts), and the patient is that which is moved (acted upon). Both the agent and the patient possess in common the will. For it is by the will that the agent acts, and it is also by the will that it is made possible for the patient to be acted upon. The agent also is moved by an end, by a good (or goal) that it intends to accomplish. The agent moves toward a goal because it understands and recognizes the goal by means of the intellect. However, it is by means of the will (appetite/desire) that the agent moves toward the goal.[4] Thus, in an agent, there must be the possession of both intellect and will. In a different sense, since the intellect moves (the person) and the will also moves (the person), both the intellect and the will could be separately said to be agents. Therefore, the human person, holistically, is an agent because of the possession of both the intellect and the will. Aquinas (ST Pt. 1, Q.82, a.4) beautifully puts the relation between the intellect and the will thus: "we can easily understand why these powers include one another in their acts because the intellect understands that the will wills, and the will wills the intellect to understand." In the metaphysics of being of the persona, it is, therefore, the powers of the intellect and the will that define an agent.

Aquinas (see ST Pt. 1, Q.82, a.1), in his analysis of different senses by which a thing could be said to be necessary, that is, the term necessity, makes distinctions that relate to action and agency. They are "natural and absolute necessity," "necessity of end," and "necessity of coercion." A natural or absolute necessity deals with the essential nature of a thing, in such a way that it deals with the definition of the thing.[5] For example, that the human being is rational is a natural necessity. For a being to be a human being, it is necessary that he or she should be rational. Hence, that agents act or produce action is a natural necessity. It is in the very nature of an agent to produce action, for whatever that lacks the capacity to produce action is not an agent. The necessity of end, strictly speaking, does not involve the direct interplay between agency and action. It deals with indispensable actions of specific ends. For example, food is necessary for life; sunlight is necessary for photosynthesis. On the other hand, the necessity of coercion strictly involves the interplay between agency (agent) and action. According to Aquinas (ST Pt. 1, Q.82, a.1): "On the part of the agent, a thing must be, when someone is forced by some agent, so that he is not able to do the contrary. This is called the necessity of coercion." In the necessity of end, there is merely an event, but in the necessity of coercion, over and above the event, there is a human subject or agency involved. Only that being with the possession of will could be said to be coerced and thus could be said to have experienced the necessity of coercion. The necessity of coercion, nevertheless, according to Aquinas, is the experience that is

repugnant to the will; it is not a natural act of the will. The human agent, in the necessity of coercion, performs actions that are contrary to the natural inclination of the will; thus, such actions are said to be not voluntary. Aquinas (ST Pt. 1, Q.82, a.1) maintains that as the intellect naturally of necessity adheres to the first principles, "the will must of necessity adhere to the last end, which is happiness: since the end is in practical matters what the principle is in speculative matters."

The notion of human action has been traditionally distinguished as voluntary or non-voluntary actions with respect to whether or not the will is involved. Contemporary philosophy of action understands the notion of action, usually by making a distinction between "mere happening" and "action," which implies the traditional distinction between non-voluntary and voluntary actions, respectively. According to Harry G. Frankfurt (1997)[6]: "The problem of action is to explicate the contrast between what an agent does and what merely happens to him, or between the bodily movements that he makes and those that occur without his making them." Employing the terminologies of the Scholastics, "mere happening," could be said to be *actus hominis* (acts of humans), which are mainly involuntary and unintentional actions. Whereas, "action" is *actus humanus* (human acts), which could be voluntary and intentional actions and/or voluntary and unintentional actions (See Donald Davidson, 1997).[7] As Karol Wojtyła (1979, 25) contends, 'It is only man's deliberate acting that we call an "act" or "action." Nothing else in his acting, nothing that is not intended and deliberate, deserves to be so termed.' [8] For Wojtyła, action deals with the subjectivity of the human person, which implies an experience of the efficacy of action in the consciousness of a person as the cause and creator of an action. He (1993, 226) contends:

> While it may be granted that the person and action—or, to put it another way, my own existing and acting self—is constituted in consciousness to the extent that consciousness always reflects the existence (*esse*) and activity (*operari*) of that self, still the experience of the human being (and especially the experience of my own self) clearly reveals that consciousness is always subjectified in the self and that its roots are always the *suppositum humanum*.

This Wojtyła's understanding of action as it relates to consciousness and a person's ontology is a very important point to note in the comprehension of the specificity of human action and agency. This is because, as will be exposed below, the naturalistic framework on which computational model in cognitive science and cognitive neuroscience is employed in the attempt to explain human action and agency makes a claim that intends not to distinguish the notion of "action" and "agency" in artificial agents from that of the human person. This understanding of human agency as the cause and creator of action makes Wojtyła conceive every human action as a moral action and the human person as a moral and/or ethical person.

Following the Aristotelian-Scholastic traditions, "action" could be distinguished

as, a. *Agere*, which implies acts and operations in general. b. *Operari*: which implies human acts with moral/ethical implications (See Karol Wojtyła, 1979). Like many terms and concepts we use in daily communication, the term "action" may appear well-understood by everyone. But the fact that there is in contemporary philosophy the special area of the discipline of philosophy called: "Philosophy of Action" shows that the notion of action is a philosophical problematic, hence, not easy to explain. There are two main subjects of philosophy of Action: a. What is action? b. How can action be explained? (Theories of Action). According to Alfred R. Mele (1997, 1), "philosophers of action want to know both what it is that explanations of actions explain and how actions are properly explained." Hence, there is an intrinsic connection between philosophy of action and philosophy of mind, as would be seen in the next paragraph. Since to define action is one of the main subject matters of philosophy of action, it is, therefore, problematic to pin down the *quiddity* of action. Nevertheless, for the purpose of this discourse, the following are working definitions of action: Action is the *quiddity* of Agency; Action is the consequence of agency; Action is the consequence of being an agent. Simply put, action, *per se*, is that which necessarily implies agency or an agent. These definitions of action, though constructed in slightly different wordings and syntax, are, strictly speaking, actually one and the same definition. This definition of action actually only states action as what is necessarily implied in agency or in an agent. Hence, it is an analytical definition of action that is more or less not saying anything new about action. In the definition: action is the consequence of agency since, at the manifestation of action, agency is immediately conceived.

In Philosophy of Action, there are various theories or explanations of action, which could be classified into these two categories: Causal and Non-causal theories of action. In explaining action, the causal theorists of action maintain that there is/are causal antecedent(s), such as beliefs, desires, intentions etc., to action, and these causal antecedents are necessary and sufficient for action to occur. Donald Davidson is one of the staunch advocates of the causal theory of action.[9] For him, every action is performed by a reason or reasons, which he calls "primary reason" (pro attitudes and beliefs). It is to this effect that he (1997, 28) states these two theses about primary reasons: 1. "For us to understand how a reason of any kind rationalizes an action, it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason." 2. "The primary reason for an action is its cause." It follows for him that the reason for an action is identical to the cause of the action. Hence, causal theory of action is grounded in the dualist theory of mind.[10] For example, a person's *belief* that a certain lecture would help him or her to know more about AI *causes* the person to sign up (action) for the lecture. In the dualist theory of mind, the mind is different from the body (brain) (especially as in substance dualism, but in property dualism, the mind is not a subsisting substance different from the body but maintains the existence of mental and physical properties as distinct), and the mental contents or phenomena are immaterial intentional qualitative experiences in the mind or consciousness. These mental states, such as beliefs, desires, wants, emotions, feelings and so on, cause a part or the whole body to move (action). On the other hand, non-causal theorists of action,

in explaining action, maintain that there is/are no causal antecedent(s) to action or maintain that, even if there is a causal antecedent, they deny that it is necessary and sufficient for action to occur. Thus, they contend that actions are simply a result of physical, chemical, and biological activities of the brain and neurons. They either dismiss the existence of mental contents or intentional states as illusions or reduce them to a neurological explanation. This gives the two main versions of materialist theory of the mind: the eliminative and reductionist theories of the mind. Hence, the non-causal theory of action is mostly grounded in the materialist theory of mind and computational theories in cognitive science. To this end, the human mind is either understood as a computer program or the brain is understood as a computer. These computational theories of the mind are employed in the different branches of cognitive science in the understanding and explanation of human action and other cognitive phenomena. The strong believers in this approach include Daniel C. Dennett (1991) and Paul M. Churchland (2013). However, it is important to note that there are non-causal theorists who are not materialists, and also there are non-dualist theorists of mind who are causal theorists of action. A good example of the latter is John Searle. John Searle (1997; 2002), even though he considers himself a naturalist, however, strongly maintains the irreducibility of the mind or consciousness to the brain. He argues that mental states and/or consciousness are to be understood as biological phenomena and can be explained (not yet, but in the future) with a better understanding of the brain and neurons. He believes that advancement in the neurosciences will bring a better understanding and explanation of the cause and nature of consciousness.

Philosophers disagree on whether the cause of an action is identical with the reason of an action. Hence, it is important to distinguish: a. Acting for a reason(s), from b. Acting because of a reason(s). Whereas acting because of a reason (belief, desire, want, intention etc.) is causal, acting for a reason, is acting not because of causal antecedents. Though both kinds of acting are closely related, but, as Robert Audi (1997)[11] argues, they are not one and the same. According to Audi (1997, 75), both are closely related in the sense that "acting for a reason is closely related to acting intentionally, to acting rationally, and to acting on the basis of practical reasoning," since, intention is a causal antecedent. The specificity of acting for a reason is acting based on or in light of a certain reason that guides the entire processes or events of an action. As Audi (1997, 80) puts it:

> Suppose *S* had believed … that dropping the breakfast tray was the best way to wake Jan, and had dropped it in order to wake her. Then *S* would have acted for, not merely because of, a reason. What has been added? When a *connecting belief*, together with a motivating want, brings about the action (in a suitable way), *S* acts not just because of but *in the light of* a reason.

It is important to emphasize the distinction between acting because of a reason and acting for a reason, however closely related they seem to be. This is, in order to

make clear the difference between action and mere happening. As Harry G. Frankfurt (1997) has argued, if necessary and sufficient explanation for action is causal antecedent(s), action and mere happening with respect to body movement may be hardly distinguished since both involve certain causal antecedents. It is as a result of this that he claims that the "causal approach is inherently implausible and that it cannot provide a satisfactory analysis of the nature of action." Hence, non-causal theorists, such as Harry G. Frankfurt, contend that, though actions may have causes (originated by causes), but an action must not necessarily have a cause or causal antecedent (causal explanation); there may, however, be a reason or reasons for the action. For him (1997), the reason for action is that which keeps an action "under the guidance" of an agent.

The discourse above has been mainly on the notion of action. Action does not produce itself. As mentioned above, for every action, there is an agent or the notion of agency. Hence below, a brief exposition of the notion of agency is discussed. Following the contemporary debates (see the work edited by Laura Waddell Ekstrom, 2001) on the question of human agency, the philosophical notion of agency could be distinguished as follows:

a. Indeterministic (incompatibilistic) Notion of Agency.
b. Deterministic Notion of Agency.
c. Compatibilistic Notion of Agency.

In order to understand what distinguishes these notions of agency, one needs to understand the bone of contention. The bone of contention has to do with the problematic of moral, ethical and/or legal responsibility, whether or not there is a necessary connection between free will and responsibility. Can one who does not act freely be said to be a free agent? As mentioned above, an agent is that which produces an action, but an agent needs to be responsible for its action, in order to be a free agent. This implies that when an agent is not responsible for his or her action, the agent would be said to be determined. Roderick M. Chisholm (2001, 126),[12] in referring to this problematic as "the metaphysical problem of human freedom," summarized it thus: "Human beings are responsible agents; but this fact appears to conflict with a deterministic view of human action (the view that every event that is involved in an act is caused by some other event); and it also appears to conflict with an indeterministic view of human action (the view that the act, or some event that is essential to the act, is not caused at all.)" Hence, the indeterministic notion of agency maintains that for there to be responsibility of any kind, the human agent needs to freely cause (as in being the source and author of) his or her action, not by any law of the natural science or historical antecedents. To have free will is a person's ability to do or act otherwise, irrespective of the consequence. In its part, the deterministic notion of agency denies free will in human agency but maintains that an action is caused as a result of certain nominal causes (as in laws of natural sciences) and/or historical antecedents (past events).[13] The compatibilistic notion of agency aims at reconciling the indeterministic and deterministic notions of agency. It maintains that determinism

is compatible with free will in human agency. The indeterministic theorists of agency, such as Peter van Inwagen (2001),[14] maintain that free will is incompatible with determinism. While John Martin Fischer (2001)[15] a compatibilist (he considers himself to be what he calls a new compatibilist), has argued that in our actions, we cannot be certain that we have alternative possibilities before us to think that we can decide freely to do otherwise.

From the ongoing, it could be observed that the deterministic notion of agency relates to the non-causal theory of action, in the sense that both share a materialistic metaphysics of the human person. For free will is a power in the human person that necessarily implies the possession of mind, (*intellectus*) the intellect. Moreover, following the Aristotelian-Scholastic tradition, since the *intellectus* is that which apprehends the intelligible forms of things, it is by nature immaterial.[16] Hence, free agency necessarily requires the possession of mind and its conscious phenomena, which are immaterial. It is as a result of this that a person can be said to possess autonomy.

## ACTION AND AGENCY IN ARTIFICIAL INTELLIGENCE

After exploring the notion of action and agency in the human person, this section exposes the notion of action and agency in artificial intelligence, otherwise called intelligent systems. Computers, especially AI, are designed and developed to do something and take action. The beginning of AI research and development is aimed at creating computing systems that can execute cognitive actions by simulating the human mind and the neurological system. In the early period, priority was given to two main cognitive actions, problem-solving and learning. While the symbolic AI or what John Haugeland (1989) calls Good Old Fashion Artificial Intelligence (GOFAI) model, which simulates the human mind, takes as its ultimate cognitive action, problem-solving, the Connectionist or Neuron Network model, that simulates the brain and neurons, takes as its ultimate cognitive action, learning. Efforts to combine and integrate these two cognitive actions in AI led to the development of the Hybrid model of both the symbolic and connectionist models. More so, in order to develop real artificial agents, current research and developments of AI aim at producing what is generally called a Situated, Embodied, and Dynamic (SED) model of AI. The SED AI model aims at developing AI systems that are not merely imprisoned in computer devices but that have the capability to interact with the environment and take action in response to information from other cognitive agents and natural phenomena in the environment as the human person does. These models in AI are also applied as models of computational research in cognitive science and cognitive neuroscience in the study of the human person and its cognitive powers. While cognitive science focuses mainly on human cognition, cognitive neuroscience, since the 1980s, has attempted to explain human body action, otherwise called motor cognition, empirically. That is to say, cognitive neuroscience attempts to utilize cognitive theories, methods, and models as well as the findings of neuroscience in the understanding and explanation of human

body action. One could also say that cognitive neuroscience is an attempt to give (materialistic/naturalistic) scientific explanation to the mind-body(brain) relation problem, which hitherto has been a problem within the bounds of philosophy of mind or psychology. It is, thus, a mechanistic explanation of how the brain causes cognitive phenomena and how the brain helps in explaining the behavioral activities of the body. As it concerns actions of the body (motor cognition), Elisabeth Pacherie (2012, 92) observes: "Work in the field of motor cognition aims at uncovering and understanding the mechanisms and processes involved in action specification and control." For example, it helps to explain whether or not there are cognitive phenomena behind, for instance, the raising of my hand, the stretching out of my leg, the blinking of my eyes and so on. It probes the question of whether or not there are causal relationships between mental phenomena (beliefs, intentions, desires, wants etc.) and the brain. A strict mechanistic stance of cognitive neuroscience maintains that the actions of the body are squarely as a result of the activities of the brain. In explaining the methodology of cognitive neurology with respect to the understanding of action, Pacherie (2012, 97) maintains that the field of motor cognition in cognitive science "integrates research techniques and methods from cognitive psychology, behavioral neuroscience, and computational modeling in an attempt to provide a unified approach to the different representations involved in the generation of action, and the contributions of different brain structures to the planning and execution of movement."

Two of the main champions of the symbolic AI model are Allen Newell and Herbert A. Simon. They (1990; 2019) posit "problem-solving" as the end of creating AI. For them, AI, as a "physical symbol system," possesses intelligence, which they understand as the symbolic representations and manipulations of symbols. In other words, for them, intelligence has to do with computation and information processing systems; anything that can process information has intelligence. This position is rooted in their (2019, 19) conviction that "human thinking and problem-solving postulates that the human operates as an information processing system." They (2019, 20) understand an information processing system (IPS) as "a system consisting of a memory containing symbol structures, a processor, effectors, and receptors." Under the influence of Newell and Simon's understanding of AI, Eduardo Alonso (2014, 232) contends: 'To get an AI system to "act" it is enough to give it a logical representation of a theory of action (how systems make decisions and act accordingly) and get it to do a bit of theorem proving.' However, Alonso (2014) observes the limitation in conceiving "acting" in AI strictly in terms of symbolic manipulations. Thus he (2014, 233) maintains: "Unfortunately, given the computational complexity of theorem proving in even very simple logics, this approach to the design and implementation of rational systems has not been widely applied in real-life scenarios." As alluded above, designing intelligent systems that can execute only theoretical (mathematical and logical) actions is not enough. The actions of reasoning or rational actions must be translated into actions that impart the environment and that, in turn, are imparted by the environment. Hence, there is a need to design alternative systems to mere symbolic systems. This leads to the design of "reactive systems." According to Alonso (2014, 233): "A reactive system is one that does not use a symbolic model of the world nor

symbolic reasoning to decide what to do next. Reactive architectures are modeled as black boxes: They follow if-then rules that directly map inputs into actions." However, Alonso (2014, 234) has this reservation about the reactive system: "Reactive systems learn procedures but no declarative knowledge; that is, they only learn values or attributes that are not easy to generalize to similar situations (or transmit to other systems)." Reactive systems do not need to symbolically represent the world. They rather act simply by following a program of if-then rules, in the form of "If this situation…… then act thus….." Examples of reactive systems are robots. Robots are designed to execute a specific action or a set of actions. They are restricted by what is generally called the relevance problem, frame problem, or the *ceteris paribus* problem—simply put, this restriction is the ability to "think outside the box," to reason and act analogically, and the ability to universalize individualized actions, ability to take action even when available data are not complete or not exactly the same.

Based on the deficiencies of both the symbolic system and the reactive system, Alonso (2014, 234) states that 'many researchers believe that in the twenty-first century for AI systems to perform "intelligently" they must be able to behave in an autonomous, flexible manner in unpredictable, dynamic, typically social domains. In other words, they believe that the "new" AI should develop agents.' Agent systems could also be called "intelligent systems." According to Aaron Sloman (1990):[17] "Incomplete information and the need to cope with long-term change in the social or physical environment require higher-order sources of action that provide learning: not only generators and comparators of motives but generators and comparators for the generators and comparators themselves." Sloman is making a case for autonomous AI agents. An agent should not only react to an environment, but more importantly, it should also be able to learn from its interactions with an environment and be able to generalize its knowledge in order to be able to improve its actions. According to Stuart Russell and Peter Norvig (2021, 37), "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." It is in this regard they (2021, 47) maintain that: "The Job of AI is to design an agent program that implements the agent function—the mapping from percepts to actions. We assume this program will run on some sort of computing device with physical sensors and actuators—we call this the agent architecture: agent = architecture + program."

To be an agent, AI needs to manifest what could be said to be the main behaviors of agents (human or not). According to Alonso (2014), the following are the main behaviors that agents should possess: Autonomous behavior, Adaptive behavior, and Social behavior. This means that any agent system should be able to make its own decisions and executes its own task; agent systems must be flexible and docile (dynamic) to learn and adapt to similar or different environments; and agent systems must be able to initiate social environment, mechanisms of interaction and cohesion, which explains the need for multi-agent systems (MAS). Among these three behaviors, it is clear that autonomy is what defines free agents, for autonomous behavior is necessary for both adaptive and social behaviors.[18] An autonomous (free) agent acts both because of a reason and for a reason. Hence an autonomous agent is the governor

or the master of not only its actions but more so of itself. When an agent acts only because of reason, it either has no autonomy or has lost its autonomy. However, it is important to note that modern scholarship on autonomous agency tends to reject the either-or conception of autonomy, which this paper maintains. Modern scholarship tends to rather posit gradation in autonomous agency. For instance, Margaret A. Boden (1996, 102) maintains: "Autonomy is not an all-or-nothing property. It has several dimensions and many gradations." She goes further to mention the three aspects of behavior or control crucial in distinguishing the gradations or dimensions of autonomy. This view of gradation in autonomy is aimed at reconstructing the notion of human autonomy that implies the Self or subjective qualitative conscious experience of personhood to accommodate artificial systems into the class of systems capable of autonomy.

## PHILOSOPHICAL CRITIQUE OF ACTION AND AGENCY IN AI

The metaphysics of action and the cognitive neurology of action tell us one fundamental fact: every single human action is a very complex phenomenon. No one theory can easily explain it. Hence, it will be difficult to computationally design intelligent systems to simulate human action in general. This is because it seems obvious, though denied by the materialistic theories of the mind-brain problem, the metaphysics of the human person is a being that is constituted of physical and spiritual or material and immaterial substances—body and mind. Hence, both the material and the immaterial substances that make up the *substratum*, the human person, are necessary for any comprehensive understanding and explanation of action in and of the human person. As Pacherie (2012, 101) contends, "actions come in various grades, from minimal, automatic, highly routinized actions to carefully preplanned actions with long-term and complex goals, and can have a psychological structure whose richness varies accordingly."

The question then is, can human action be explained exclusively mechanistically, that is, purely by natural science? This question presupposes the affirmation of the human person as a machine. It is, therefore, a conviction that the human person is a machine and the human mind/brain is a computer, which informs the analogical explanation of the human person and computer, which inspires the design and development of AI. Hence, notions that reference the human person *in proprio*, such as "intelligence," "reasoning," "learning," "action," "agency," "autonomy," and so on, are used to reference AI. In most cases, computer scientists, AI researchers and developers, and even cognitive scientists employ these notions in their description and explanation of AI in the manner that these notions are used in describing the human person. This is definitely misleading. It is the epistemic responsibility of philosophy to engage in a critical analysis of these notions by expounding the *quiddity* of these notions in order to give a rigorous and profound critique of AI. This is the reason for expounding above the metaphysical notions of "action" and "agency" before the brief exposition of these notions in AI in order to construct a philosophical critique of these

notions in AI.

The fact that AI is designed and developed to do something simply means that AI performs actions. So, there is the notion of "action" in AI. The question, however, is: Is the notion of "action" in AI the same as the notion of "action" in the human person? Philosophy of Action and Cognitive neuroscience, as briefly exposed above, tell us that human action is not only very complex to explain but that there is yet no comprehensive explanation of human action, even as simple as giving an explanation for the body's action of blinking the eyes or raising a hand. Moral, ethical and legal actions by the human person are even more complicated to explain. These complications and complexities inherent in the understanding and explanation of human actions are because of the notion of "agency" in the human person. Most philosophers of action are of the view that the notion of "agency" necessarily entails the notion of "free will." As exposed above, some theorists, the determinists/compatibilists, deny this position. They either contend that "agency" does not necessarily entail "free will," or they deny free will while affirming agency in the human person. This raises the question of moral, ethical and/or legal responsibility.

For there to be moral, ethical, or legal responsibility, a place for blame or praiseworthiness in actions, an agent must be rational and free. Rationality is important to know and understand an act and its consequences, at least to a certain degree, for an agent to be held responsible for an act. No one holds an animal, a human baby, or even a human adult with mental problems morally, ethically and/or legally responsible. Simply, the reason has to do with a judgment on their lack of rationality completely (or insufficiently present) for the understanding of an act. Even for a rational, healthy adult human person, he or she must be judged to act freely in order to be responsible for an action. This acting freely goes with not only having a clear and culpable knowledge of an action, but even more so, the person has *to intend to* do the action before he or she can be judged to be responsible for the action. Therefore, the question is: If we concede that AIs, are rational, intelligent systems, can we also concede that AIs are free (possessing free will)? The notion of "intelligence" in AI, as symbolic representation and manipulation as expounded by Newell and Simon, has been exposed above. Even so, this does not qualify as the notion of "intelligence" in the human person, following the Aristotelian-Scholastic philosophical tradition. For the Aristotelian-Scholastic tradition, "intelligence" is more than a mere symbolic representation and manipulation, for the notion of "intellectus" is that which makes the apprehension of the intelligible forms of things possible. Intelligence, therefore, in the human person, is the comprehension of the intelligible forms (the *quiddity*) of things (including actions), not merely a symbolic representation and manipulation of things in the natural world, which are merely phantoms in the mind, according to the Aristotelian-Scholastic tradition. For instance, an action can be symbolically represented, but the moral/ethical good of an action or its contrary cannot be merely symbolically represented. It is rather, directly apprehended by the human intellect.

Nevertheless, since AI executes certain main characteristics of the possession of intelligence: reasoning, decision-making, learning, and problem-solving, one can concede that AI has a certain level of rationality but not (and this author agrees with

philosophers like John Searle (1990;1997) and Hubert Dreyfus (1986;1992), that AIs cannot now or in future have) the same rationality as in humans. It follows that AI can execute certain actions, and it can execute certain intelligent actions, of which some of these actions are executed even more efficaciously than the human person. However, AI cannot execute free actions (moral or ethical) for these two main reasons: AI is not a free agent, and AI cannot apprehend the moral good of actions. Ordinary computers cannot be said to have actions but mere happenings. Thus, the need to design and develop AI. As affirmed above, AI can produce actions, but not free actions. Hence, since an AI system can produce action, it could be said to be an agent. But since AI cannot produce free actions, it is not a free agent. Therefore, AIs are deterministic agents (do not have free agency), while the human person only has free agency. Being a deterministic agent implies that it does not have autonomy; that is to say, it does not have causal efficacy—it is not both the agent and the creator of its actions. Hence AI does not produce *operari*—actions with moral and ethical implications (being responsible, being praised or blamed) - having the self-awareness of "I ought to act thus…but I will act otherwise." As Peter van Inwagen (1975) contends: "Almost all philosophers agree that a necessary condition for holding an agent responsible for an act is believing that that agent could have refrained from performing that act." Hence, free will is a unique power in the human person that is necessarily connected to the capability to grasp the transcendental properties: The One, the Good, the Truth, and the Just. These transcendental properties are the fundamental grounds for the knowledge, understanding and judgment of the moral and ethical nature of actions. AI cannot grasp these transcendental properties. It cannot because they are metaphysical properties that cannot be mathematized nor *logicized*; hence, they are not and cannot be computable. No computer scientist can write a program or algorithm of the good, the truth, or the just. For even within philosophy, that could be said to be the epistemic scope of these properties; these properties remain members of the perennial problems in philosophy.

In summary, the overall argument in this paper is thus:

P1: AI can perform certain intelligent actions (from experience).
P2: AI is capable of actions (based on P1).
P3: To be capable of performing actions implies agency (from the definition of agency).
P4: AIs could be said to have agency (from P2 and P3).
P5: Not all actions are free actions (even in the human person, from experience).
P6: It takes not only rationality but, most importantly free-will to perform free actions (even in the human person).
P7: AI does not have free-will (if one does not hold a materialistic-deterministic-non-causal view of the person and action).
P8: AI cannot perform free actions (based on P6 and P7).
P9: AI is not a free agent (from P3, P6, P7, and P8).
P10: Free agency implies autonomy (from the definition).
P11: AI cannot have autonomy (from P9).

P12: Autonomy is necessary for moral and ethical responsibility (by definition, autonomy is non-deterministic, giving room for moral and ethical responsibility).
P13: Therefore, AI cannot be morally and ethically responsible (from P12).

Therefore, as the research and design of AI advances and AI systems become cohabitant with humans, as closer associates to humans in our day-to-day existential experiences, this philosophical reflection on action and agency in artificial intelligence calls for a socio-political response to the question of moral, ethical and/or legal responsibility in AI. Since AI, *per se*, is not a free (moral and ethical) agent, as argued in this paper, it is then the duty of individual nations, or the global community to define and enact policies on who shoulders the responsibility of actions executed by AIs.

## CONCLUSION

In conclusion, this paper has exposed the notion of "action" and "agency" in artificial intelligence, by expounding the metaphysics of these notions as they relate to the human person. It has explored works in Philosophy of Artificial Intelligence, Action, Mind, and Consciousness, and also works in Cognitive Science and Cognitive Neuroscience in order to give a broad and systematic analysis of the notion of "action" and "agency." A philosophical critique on these notions in AI has been expounded to clarify the limit by which these notions could be or cannot be used with respect to artificial intelligence. Finally, a socio-political recommendation has been given with respect to the question of responsibility in the actions executed by artificial intelligence.

## NOTES

1. There is a difference between moral and ethical actions; however, close they may appear in general usage. An action could be ethical but not moral. An ethical action deals with right and wrong actions for a cohesive and harmonious relationship by a group of persons. So, it takes at least two persons for an ethical action to be in place. Moral actions deal with good and evil (bad) actions. So, a person's action could be judged to be good or evil, even when the action does not affect another person or the society directly. So, moral actions question the cultivation of virtue in a person. If I refuse to share my bread with a person who is famished, even if my action may not be judged to be ethically wrong since I am not obliged to be generous by any ethical prescription, however by that action, I will be judged to be a morally bad or evil person. For it questions the virtue of kindness and generosity in me.

2. The Latin term *quiddity* is used a number of times in this paper. In Scholastic metaphysics, it has a meaning which is different from the term "nature" in general. The term "nature," when used generally, could imply the term *quiddity*, but strictly speaking, both are not the same. *Quiddity* means that which makes a thing that thing, that is to say, the essence (*ousia*) of the thing. As the scholastics will say, every

definition has species and genus; the *quiddity* of a thing is the species in the definition of a thing. For instance, in the definition: A human person is a rational animal. "Rational" is the species, while "animal" is the genus. It means that to be "rational" or rationality is the *quiddity* of the human person. However, the "nature" of the human person, strictly speaking, is not only what defines the human person; it includes other metaphysical specificities that define the human person. For instance, speech, to walk with two legs, and to be able to grasps objects with fingers are all in the nature of the human person. However, they are also in the nature of some other animals (primates). Thus, one can still conceive the human person in the absence (privation or deficiency) of these natures of the human person, but one cannot conceive the human person without implying rationality. Hence, by investigating the metaphysics of the notion of "action" and "agency," the intention is to elucidate what makes action and agency per se, not merely as in the senses by which they are employed by linguistic conventions or as technical times in the profession of AI or cognitive science.

3. According to Aquinas in *Summa Theologica* (ST), whereas the intellect is superior to the will in the absolute sense of superiority of one thing over another when they are considered in respect to themselves. Since he (ST Pt 1, Q,82, a.3) argues, "the object of the intellect is more simple and more absolute than the object of the will." However, he (ST Pt. 1, Q.82, a.3) maintains that "relatively and by comparison with something else, we find that the will is sometimes higher than the intellect, from the fact that the object of the will occurs in something higher than that in which occurs the object of the intellect."

4. See Aquinas (ST Pt. 1, Q,82, a.4) for more understanding of the relations between the intellect and the will as agents.

5. It is in this sense of necessity, the author thinks, that James Ross (2008, 18), though in distinguishing "natural necessities" from "formal necessities" (not in respect to action and agency but in respect to truth), maintains: "Necessities of nature such as "humans can think" earn truth by expressing what is naturally so, and have necessity by being so "no matter what."

6. Harry G. Frankfurt's essay, "The Problem of action," was originally published in 1978. The reference here is from its re-publication in the collection edited by Alfred R. Mele (1997).

7. The reference is to Donald Davidson's essay "Actions, Reasons, and Causes," originally published in 1963, and re-published in the collection edited by Alfred R. Mele (1997). He argues that an action can be done voluntarily with or without a (primary) reason, that is, without a purpose. To act intentionally is to act with a purpose, to act directed towards an end. As he argues (on p.31): "To know a primary reason why someone acted as he did is to know an intention with which the action was done…. But to know the intention is not necessarily to know the primary reason in full detail." For the sake of elucidation, let us assume this scenario wherein I throw a pen to someone (voluntary) because she needs a pen to write, but the pen hits one of her eyes without me intending to injure her eye. This action that leads to causing injury to her eye, though voluntary, is not intentional.

8. Since this action is what, for Wojtyła (1993), that which truly defines the

human person, he thinks that it should be more accurately called *actus personae* (act of person).

9. Other philosophers of action that have argued for the causal theory of action include Jennifer Hornsby in her essay entitled "Agency and Causal Explanation," first published in 1993 and re-published in Alfred R. Mele (1997).

10. It is important to note that the question here is not whether Davidson is a "property dualist" or "substance (Cartesian) dualist." Davidson's theory of action is only referenced as a causal theory of action to make the claim that the "causal theory of action is grounded in the dualist theory of mind." So, it is immaterial, whether it is substance or property dualism. For if one holds that mental states have a causal relationship with actions, it is clear that the person is a dualist. However, whether one is a substance or property dualist is a different matter.

11. Robert Audi's essay "Acting For Reasons" was originally published in 1986. The reference here is from its re-publication in the collection edited by Alfred R. Mele (1997). The distinction between acting as acting because of a reason and acting for a reason is credited to Audi, who employed them in this essay.

12. This Roderick M. Chisholm's essay "Human Freedom and the Self" was originally published in 1964. The reference here is from its re-publication in the collection edited by Laura Waddell Ekstrom (2001).

13. Some philosophers maintain the position of what is called "soft determinism." David Lewis, in his essay, "Are We Free to Break the Laws?" was originally published in 1981. The reference here is from its re-publication in the collection edited by Laura Waddell Ekstrom (2001, 30), which defines it thus: "Soft determinism is the doctrine that sometimes one freely does what one is predetermined to do; and that in such a case one is able to act otherwise though past history and the laws of nature determine that one will not act otherwise." He tries to correlate soft determinism with compatibilism by maintaining that: "Compatibilism is the doctrine that soft determinism may be true. A compatibilist might well doubt soft determinism because he doubts on physical grounds that we are ever predetermined to act as we do, or perhaps because he doubts on psychoanalytic grounds that we ever act freely." It is due to this correlation and distinction that he considers himself to be a non-deterministic compatibilist.

14. For a clear argument for indeterminism/incompatibility, see Peter van Inwagen's essay "The Incompatibility of Free Will and Determinism," originally published in 1975. The reference here is from its re-publication in the collection edited by Laura Waddell Ekstrom (2001).

15. John Martin Fischer, in his essay "A New Compatibilism," originally published in 1996, makes a strong case for compatibilism. The reference here is from its re-publication in the collection edited by Laura Waddell Ekstrom (2001).

16. For a comprehensive discourse on the notion of *intellectus,* see the following: Moses Maimonides (2017,100-102); Thomas Aquinas (ST Pt. 1, Q75-Q79).

17. Aaron Slomon's essay "Motives, Mechanisms, and Emotions" was originally published in 1987. The reference here is from its re-publication in the collection edited by Boden A. Margaret (1990).

18. Some contemporary scholars maintain degrees or dimensions of autonomy. This explains the reason to explore the metaphysics of action and agency. It is not the intention of this work to expose the conventional view of autonomy and the usage of the term autonomy in AI, in particular by AI researchers. By a metaphysical exposition of action and agency, a thing could be said to either have autonomy or not. So, I argue here that by autonomy *per se*, AIs, since they are not free agents, cannot be said to have autonomy. Hence, when the term "autonomy" is used with respect to AI, it is used simply metaphorically. So, in my line of thought, I do not accept "degrees" of autonomy as "non-autonomous," "semi-autonomous," and fully autonomous" agents, as suggested by some scholars.

## REFERENCES

Alonso, Eduardo. 2014. "Actions and agents" in *The Cambridge handbook of artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Cambridge: Cambridge University Press, pp. 232-246.

Aquinas, Thomas. 1948. *Summa theologica*, Translated by Fathers of the English Dominican Province, New York: Benziger Brothers.

Audi, Robert. 1997. "Acting for reasons", in *The Philosophy of action*, edited by Alfred R. Mele, New York: Oxford University Press, pp. 75-105.

Boden, Margaret A. 1996. "Autonomy and artificiality," in *The Philosophy of artificial life*, edited by Margaret A. Boden, Oxford: Oxford University Press, pp. 95-108.

Chisholm, Roderick M. 2001. "Human freedom and the self," in *Agency and responsibility: Essays on the metaphysics of freedom*, edited by Laura Waddell Ekstrom, Colorado: Westview Press, pp. 126-137.

Churchland, Paul M. 2013. *Matter and consciousness*, Cambridge: A Bradford Book/The MIT Press.

Davidson, Donald. 1997. "Actions, reasons, and causes," in *The Philosophy of action*, edited by Alfred R. Mele, New York: Oxford University Press, pp. 27-41.

Dennett, Daniel C. 1991. *Consciousness explained*, New York: Back Bay Books.

Dreyfus, Hubert L. & Dreyfus, Stuart E. 1986. *Mind over machine: The Power of human intuition and expertise in the era of the computer*, New York: The Free Press.

Dreyfus Hubert L.1992. *What computers still can't do: A Critique of artificial reason*, Cambridge: The MIT Press.

Fischer, John Martin. 2001. "A New compatibilism" in *Agency and responsibility: Essays on the metaphysics of freedom*, edited by Laura Waddell Ekstrom, Colorado: Westview Press, pp. 38-56.

Frankfurt, Harry G. 1997. "The Problem of action," in *The Philosophy of action*, edited by Alfred R. Mele, New York: Oxford University Press, pp. 42-52.

Hornsby, Jennifer. 1997. "Agency and causal explanation," in *The Philosophy of action*, edited by Alfred R. Mele, New York: Oxford University Press, pp. 283-307.

Inwagen, Peter van. 2001. "The Incompatibility of free will and determinism" in

*Agency and responsibility: Essays on the metaphysics of freedom*, edited by Laura Waddell Ekstrom, Colorado: Westview Press, pp.17-29.

Lewis, David. 2001. "Are we free to break the laws?" in *Agency and responsibility: Essays on the metaphysics of freedom*, edited by Laura Waddell Ekstrom, Colorado: Westview Press, pp. 30-37.

Maimonides, Moses. 2017. *The Guide for the perplexed*, Translated by M. Friedländer, New York: Dover Publications, Inc.

Newell, Allen and Simon, Herbert A. 1990. Computer science as empirical enquiry: Symbols and search. In *The Philosophy of artificial intelligence*. Edited by Boden A. Margaret. Oxford: Oxford University Press, pp. 105-132.

Newell, Allen and Simon, Herbert A. 2019. *Human problem-solving*. Vermont: Echo Point Books & Media.

Pacherie, Elisabeth. 2012. "Action," in *The Cambridge handbook of cognitive science*, edited by Keith Frankish and William M. Ramsey, Cambridge: Cambridge University Press, pp. 92-111.

Ross, James. 2008. *Thought and world*, Indiana: University of Notre Dame.

Russell, Stuart and Norvig, Peter. 2021. *Artificial intelligence: A Modern approach*, New Jersey: Pearson Education, Inc.

Searle, John R. 1990. "Minds, brains, and programs," in *The Philosophy of artificial intelligence*, edited by Boden A. Margaret, Oxford: Oxford University Press, pp. 67-88.

Searle, John R. 1997. *The Mystery of consciousness*, New York: The New York Review of Books.

Sloman, Aaron. 1990. "Motives, mechanisms, and emotions," in *The Philosophy of artificial intelligence*, edited by Boden A. Margaret, Oxford: Oxford University Press, pp.231-247.

Wojtyła, Karol. 1979. *The Acting person*, Andrzej Potocki (trans.), Dordrecht: D. Reidel Publishing Company.

Wojtyła, Karol. 1993. "The Person: Subject and community," in *Person and community: Selected essays of Karol Wojtyła*, pp. 219-261.