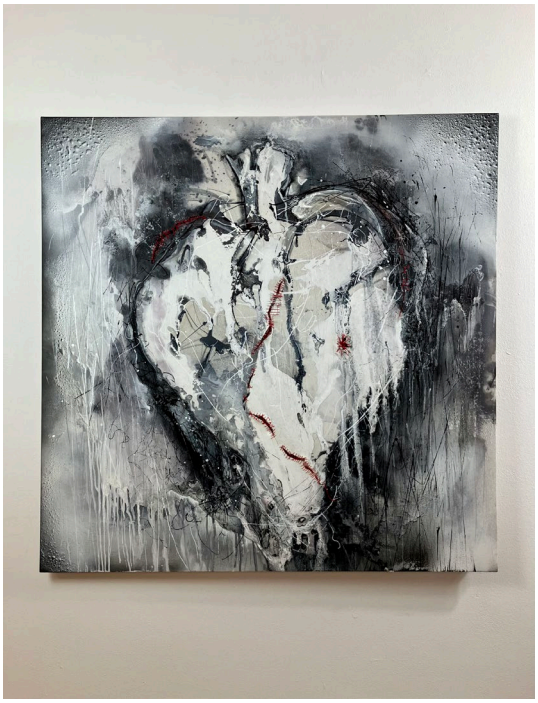
FRANCESCO STRIANO

CAN ARTIFICIAL AGENTS ACT?
CONCEPTUAL CONSTELLATION FOR A DE-HUMANISED THEORY OF ACTION[1]

ABSTRACT: ***CAN ARTIFICIAL AGENTS ACT? CONCEPTUAL CONSTELLATION FOR A DE-HUMANISED THEORY OF ACTION***
This paper embarks on an exploration of the concept of agency, traditionally ascribed to humans, in the context of artificial intelligence (AI). In the first two sections, it challenges the conventional dichotomy of human agency and non- human instrumentality, arguing that advancements in technology have blurred these boundaries. In the third section, the paper introduces the reader to the philosophical perspective of new materialism, which assigns causal power to matter itself. This perspective suggests that agency is an emergent property of material configurations, prompting a re-evaluation of nonhuman agency. The fourth and fifth section revisit the legacy of cybernetics to understand systemic properties and feedback mechanisms, while re-admitting in the discourse also linear conditioning (discarded by new materialism) and assigning it a role in system dynamics. In the sixth section, in the light of the conceptual background examined so far, the paper proposes a revision of determinism (again partly in opposition to the new materialism and its indeterministic view) that can include both linear conditioning and circular interactions. The seventh section is devoted to propose a novel theory of action that includes AI systems – and artificial entities in general – as agents that can impact their environment and human systems. The exploration concludes with a discussion on the implications of this perspective for our understanding of action and responsibility in the age of AI.



1. *Premise*

Can we define artificial intelligences (AIs) as "agents"? Are they, as we philosophers would say, endowed with *agency*? There are people who not only answer this question in the affirmative, but even argue that we should no longer speak of "intelligences," but replace the term with "artificial agents." These include well-

[1] A first version of this paper was presented and discussed during a seminar of researchers affiliated to the Chair of Moral Philosophy at the University of Turin. I would like to thank all the participants – in particular Norberto Albano, Cristiano Calì, Matteo Cresti, Laura Gorrieri, Accursio Graffeo, Paolo Monti, Andrea Osti, Marco Pavanini, and Giacomo Pezzano – for their feedback, which has been incorporated into the final draft.

known names from research and development, such as Stuart Russell and Peter Norvig[2].

For those who understand agency as the ability to act purposefully (and successfully – or, at least, potentially successfully), without heteronomy and in a way that is not strictly mechanical and linear, there is no doubt that an AI can be defined as an agent and that, compared to other applications of analogue or digital technologies, it has emancipated itself from the need to follow pre-programmed instructions and is capable of finding new solutions to problems that are similar, but not identical, to those for which it was trained. For technology enthusiasts, this is proof that neural networks can indeed be described as intelligent. For Luciano Floridi, on the other hand, it is indeed "revolutionary," but the revolution consists in having «decoupled the ability to act successfully from the need to be intelligent»[3]: a machine cannot «understand, reflect, consider or grasp anything»[4], but that is not necessary to act either.

One could argue that plants are also capable of acting purposefully and successfully, although historically there has been a tendency to attribute intelligence to the animal kingdom rather than the plant kingdom. However, there are certainly people who are willing to talk about the intelligence of plants[5] actually being able to learn from the environment, defend themselves and solve problems. Whether this is synonymous with intelligence or whether it also means "understanding, thinking, reasoning or grasping" probably depends on the definition of intelligence we

[2] See C. Calì, *Come ci cambia la tecnologia. L'Agency delle AI e la capacità cognitiva di prendere decisioni razionali*, in «S&F_scienzaefilosofia.it», 30, 2023, pp. 366-385: 366.

[3] L. Floridi, *AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models*, in «Philosophy & Technology», XXXVI, 15, 2023, pp. 1-7: 5.

[4] *Ibid.*, pp. 5-6.

[5] The Italian botanist Stefano Mancuso is emblematic in this respect. Among his numerous popular works on this subject, see, for example, S. Mancuso, *The Revolutionary Genius of Plants: A New Understanding of Plant Intelligence and Behavior*, Atria Books, New York 2018.

buy. But this only confirms how slippery the question of intelligence is and how tempting the idea of talking about AAs (Artificial Agents) instead of AIs is.

At this point a first doubt might arise: is the capacity for *purposeful behaviour* really enough to define agency? This paper will ultimately argue that, yes, this is an appropriate definition, provided it is framed within a certain theoretical horizon. Yet it may not be so obvious.

Traditionally, agency was viewed as an inherently human attribute. In this conventional view, actions were the exclusive purview of conscious beings, with the autonomy to assert their will upon the world. However, the rapid advancement of technology has ushered in an era where complex artificial entities, from artificial intelligences with machine learning capabilities to robotic systems, and even seemingly mundane machines (media machines above all), exert tangible effects on the world, on us, and on each other.

In this shifting landscape, we navigate uncharted territories, where traditional boundaries between human agency and non-human instrumentality blur. Herein lies the impetus for a recalibration of our intellectual compass.

2. *Agency and responsibility*

The concept of agent, on the other hand, seems to be no less problematic than that of intelligence. What is an agent? What does it mean to act? What is purposeful behaviour and under what conditions is it attributed to an agent? Do will or awareness play a role, or are these just human levels of explanation for something that can be explained by tracing it back to lower-level processes?

As we can see, the issue is already complicated, and the last two questions in particular point to a moral problem of no small importance: that of responsibility. In the case of artificial

intelligence[6], of decision-making algorithms trained with machine learning (ML), of actions carried out by automated systems, to whom is the origin of the purposeful behaviour to be attributed? Can the responsibility be attributed to the machine, or should it be attributed to those who build or train it? And in the latter case, should we say that the machine's agency – if it has any – is a kind of derived agency that is always traceable to human agency? These problems are already urgent if we consider that decision-making algorithms are already being used in some countries in the judiciary or public administration. The recent scandal denounced in the parliamentary report *Ongekend onrecht* ("Unprecedented injustice"), which shook the Netherlands and forced the government to resign, is a case in point: more than 26,000 families were falsely accused of fraud by the Dutch tax authority. The basis for the accusations of tax fraud was the decision of the SyRi (System Risk Indicator) algorithm. The undesirable effects of this and other algorithms often also have a racist flavour due to biases in the data sets on which these algorithms are trained. But also consider that today it is possible to attend a meeting of a company's board of directors in which humans, who are able to explain the basis of their decision-making processes, and algorithms, which, while not providing explanations of the processes underlying their judgements – which perhaps elude their own programmers – can nevertheless make decisions, can sit at the same time.

And if on the one hand it seems paradoxical for a court to "condemn" an algorithm, on the other hand it seems unconvincing to argue that those responsible for the decision or action of an algorithm are always the builders, programmers, designers, or

[6] As we shall see, my aim is to guarantee agency to any artificial entity, including any form of artificial intelligence. However, the most striking cases of possible unforeseen and unpredictable AI-related behaviour, which pose significant problems from the point of view of assigning responsibility, mainly concern AI from machine learning onwards (in particular neural networks and deep learning).

trainers. To say that those who built or supervised the learning of a machine are responsible for its subsequent behaviour – or at least for its not immediately foreseeable consequences – would be like saying that my parents or my school teachers are responsible for what I am currently writing.

So, can algorithms and AIs be held accountable? And does this mean that they are endowed with agency? To answer this question, we must first try to better define what agency is, or at least take a position on it. Then we must determine whether a certain definition of agency can be used to affirm the existence of non-human, perhaps even non-animal, or non-living, even artificial agents.

3. *New Materialism*

The standard conception of agency views it as the ability to act intentionally stemming from an agent's mental states. However, this raises questions about the exact nature of these states and their role in causing actions. Some critics argue that agency extends beyond intentionality and can be spontaneous or a manifestation of will[7]. This view does not rely on mental states but still involves a sort of "manifestation" of – more or less individual – will, making the definition of agency unclear.

In the first case, moreover, agency seems to be something that an agent possesses; in the second case, it seems that agency is something that creates agents. However, to justify the existence of artificial agents, of particular non-biological configurations that appear to be capable of purposeful and successful action, we need a different conception of agency and a new theory of action that follows from it. Among the non-standard approaches, new materialism is the one that offers a perspective where causal

[7] For an effective summary of standard and non-standard positions, as well as an exhaustive bibliography on the subject, see the entry M. Schlosser, *Agency*, in E. N. Zalta (a cura di), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), https://plato.stanford.edu/archives/win2019/entries/agency/ (last accessed 15/01/2024).

power lies in matter, and agency emerges from material configurations.

New materialism sees matter as self-transforming and self-organizing[8], prompting a re-evaluation of nonhuman agency. It proposes concepts like "distributed agency" or "agency of assemblage" between human and nonhuman elements[9]. Instead of focusing on the causal power of individuals, whether human or not, new materialism looks at *events*, considering individuals as components of these events.

Karen Barad introduces the concept of *intra-action*, a form of causal interaction where «part of the universe mak[es] itself intelligible to another part in its ongoing differentiating intelligibility and materialization»[10]. This leads to a definition of agency as «the relationality […] that and by which matter and things are defined, distributed, and organised»[11]. This conception avoids both constructivism and idealism[12], allowing us to recognize an agential role for matter from the perspective of «an ongoing topological dynamics that enfolds the spacetime manifold upon itself»[13].

This view of agency and causality rejects both technodeterminism and free will theory, seeing intra-actions as «constraining but not determining»[14]. It suggests an open future where agency is not

[8] See D. Coole, S. Frost, *Introducing the New Materialism*, in *New Materialism: Ontology, Agency, and Politics*, edited by D. Coole, S. Frost, Duke University Press, Durham-London 2010, pp. 1-43, p. 10.

[9] See B. Bargetz, *Longing for agency: New materialisms' wrestling with despair*, in «European Journal of Women's Studies», XXVI, 2, 2019, pp. 181-194, p. 187. In this passage Bargetz quotes and comments on J. Bennett, *Vibrant Matter: A Political Ecology of Things*, Duke University Press, Durham-London 2010. Jane Bennett refers to Gilles Deleuze and Félix Guattari in defining her notion of "distributed agency", but I invite to consider also G. Simondon, *On the Mode of Existence of Technical Objects*, Univocal Publishing, Minneapolis 2017, pp. 53 and ff.

[10] K. Barad, *Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter*, in «Signs», XXVIII, 3, 2003, pp. 801-831, p. 824.

[11] F. J. Colman, *Agency*, available at: https://newmaterialism.eu/almanac/a/agency.html (last accessed 15/01/2024).

[12] K. Barad, *op. cit.*, pp. 824-825.

[13] *Ibid.*, p. 826.

[14] *Ibid.*

something possessed, but «something that happens»[15], with the future remaining radically open.

4. *Two levels of explanation*

New materialism's concept of a radically open future and material agency as an event offers a useful framework for explaining unpredictable AI outcomes without attributing them to intelligence or consciousness. However, defining agency solely as a systemic property raises issues.

Linear conditioning exists in systems[16], suggesting a directed causal power of one component over another. Machines can condition human behaviour linearly, as seen in assembly lines or PC games that "trains" the worker or user to act according to the machine's own rhythms, even conditioning their body and posture. While avoiding technodeterminism is crucial, it is also dangerous to deny any causal power to system components.

Striving for an open future risks discarding essential tools for achieving this goal. It is challenging to consider action efficacy and transformative capacity without considering actors, causal power, and a form of determinism. Avoiding techno-determinism may lead to techno-fatalism, where technology outcomes are uncontrollable events.

New materialism's conception of agency should not be entirely discarded. It recognizes matter's agential role and offers a perspective that separates agency from intelligence or intentionality, making the decoupling noted by Floridi less surprising or revolutionary. The concept of intra-action shifts the focus from individuals to a systemic level.

To explain all systemic changes and ensure purposeful action *within* the system, it is nevertheless necessary to recognize that

[15] B. Bargetz, *op. cit.*, p. 188.

[16] I define "system" not as a simple collection of individuals and elements, but as an organic totality of relationships, each interconnected and all governed by laws that emerges as properties of the system itself.

system components have unique properties unexplainable by *indifferent* matter alone. Classifications at the *interactive* kinds[17] level are needed. The explanation of systemic totality provided by intra-action should be complemented by the explanation of systemic plurality provided by *interaction*.

In this dual intra-inter-active perspective agency is certainly not a property *inherent in* agents, nor is it a mysterious force *that produces* agents, but neither is it correct to define it as "something that appears". Agency is *something that is done*, in the sense that it appears through doing.

In order to understand the properties of the interaction and the form of determinism that confers causal power to system parts, we must refer to first-order cybernetics.

5. *Cybernetic legacy*

Cybernetics is *a general theory of machines*[18]. Or, in the words of Wiener, cybernetics is *a science of control and self-regulation* that enables *a unified study of natural and artificial systems*. The mathematical background is obvious, because the basic idea is

[17] The distinction between indifferent and interactive kinds was established by I. Hacking, *The Social Construction of What?*, Harvard University Press, Cambridge (MA)-London 1999. An indifferent kind is a class of individuals that is not affected by the way we classify it: «The classification ‘‘quark’’ is indifferent in the sense that calling a quark a quark makes no difference to the quark» (*ibid.*, p. 105). Classic examples of indifferent kinds are the objects of the physical sciences, but also the very concept of "matter." This does not contradict the new materialism, since «[i]ndifferent does not imply passive. The classification plutonium is indifferent, but plutonium is singularly nonpassive. It kills» (*ibid.*). Hacking instead links the concept of interactive kinds to consciousness and self-consciousness (*ibid.*, 103-104), but I argue that this is not necessary: the most important thing that defines an interactive kind is that the individuals that belong to it also change or modify their agency as a group, depending on how they are described and treated and how they interact. In the machinic realm, a transistor (technical element) belongs to an indifferent kind, since the description of the transistor does not change its function. A computer (technical individual), on the other hand, can be a tool, an agent, a component of a global network, a node of information flows, an experimental or entertainment medium, etc., depending on its use and the narratives and descriptions surrounding it.

[18] "[N]ot merely a theory of the machines that had been built already, but a theory of all machines, including those that had not been invented yet" (T. Rid, *Rise of the Machines. The Lost History of Cybernetics*, Scribe, Melbourne-London 2016, p. 4).

that natural systems or artificial machines can be symbolised and studied through models. But cybernetics is not just mathematics: it was, especially in its beginnings, an extensive interdisciplinary research programme involving engineering, biology, humanities, and social sciences. The unity of this systematic approach lies in *two concepts* taken from the philosophy of the one who is defined by Wiener as «the patron saint for cybernetics»[19], namely Leibniz: (i) universal symbolism and (ii) calculus of reasoning.

Wiener's interest in a common symbolisation of systems arose from the study of machines capable of simulating human activity, as well as from his collaboration with Vannevar Bush on analogue computers or with John von Neumann, who worked on the design of the first fully digital computer. However, to conceptualise a unified investigation of systems, it is not enough to establish a mere *analogy* between machines and living beings. There must be elements and mechanisms *common to any organisation that can be identified as a system*, be it natural or artificial.

These common elements and mechanisms can be explained using key concepts such as *system*, *information*, *metastability*, *modulators*, *feedback*, *control*, and *communication*.

By *system* is meant an organised totality: it is a set of elements and components that are relatively autonomous with respect to some properties, but each of which is dependent on and connected to the other components; this set also has its own properties, which can only be observed when it is considered in its totality; finally, it has a dynamic nature and fixity represents its death: it has a relatively stable (or metastable) equilibrium, but it is based on an internal dynamic tension.

An important common element of natural and artificial systems is *information*. Wiener's mathematical definition of information is

[19] N. Wiener, *Cybernetics. Or Control and Communication in the Animal and the Machine*, The MIT Press, Cambridge (MA) 1985, p. 12.

based on a suggestion by von Neumann and, in its simplicity and generality, can describe the basic behaviour of any agent within a system, whether natural or artificial: «One of the simplest, most unitary forms of information is the recording of a choice between two equally probable simple alternatives, one or the other of which is bound to happen»[20].

For continuous communication or propagation of a form – i.e., information –, a system must avoid a stable equilibrium and each choice should generate a new bifurcation. For this reason, Simondon introduces the concept of information tension, which is maximum when a system nears contradiction without contradicting[21]. When information tension reaches its peak, we have a *metastable equilibrium*. *Modulators* maintain this tension, facilitating communication between system parts and allowing self-regulation[22]. Particularly, modulators between machines and environments – i.e., *interfaces* – promote the emergence of systemic properties.

The most important emerging property of systems is *feedback*. The concept of feedback was introduced and illustrated by Wiener[23] as the basis for the mechanism of information circulation in biological systems, but also in electrical or mechanical systems. In 1948, Wiener cited the signal tower, the thermostat, or Maxwell's governor of a steam engine as examples[24]. In all these cases, sensors or transducers are involved that collect data and translate them into information that regulates the operation of the machine. In other words, feedback is the mechanism by which

[20] N. Wiener, *op. cit.*, p. 61.
[21] See G. Simondon, *Individuation in Light of Notions of Form and Information*, University of Minnesota Press, Minneapolis-London 2020, p. 688.
[22] A fairly simple example applied to artificial systems is that of sensors. An optical sensor, for example, is a photoconductive device that measures a change in incident light in the form of changes in resistance. It is what Simondon would call an "element" and, when inserted into a more complex system, can have various functions, such as activating a relay that switches on a light or opens a door when the natural or, as in the case of an infrared system, artificial light source is obscured by the passage of a body.
[23] See N. Wiener, *op. cit.*, pp. 95-96.
[24] *Ibid.*, p. 97.

the result of a system's action is reflected back to the system in order to correct, change, or reinforce its behaviour.

The concept of feedback is the «technoepistemic core of cybernetic (self)-governance»[25], for it is the mechanism that enables the level of interaction to communicate with the level of intra-action: the interaction and mutual control between the parts enables the self-regulation of the totality.

In an ideal condition, the *control* between the components is always reciprocal and forms the basis for the self-regulation of the system, human-machine systems included. From the machine's point of view, the *interface* – a kind of modulator – serves to focus the human component on the flow of information that the machine needs and to train it to use it correctly, which would consist of constantly providing the machine with inputs. From the human perspective, the interface is that which, while cutting out a potential part of the perceptible, opens up new possibilities for action and also allows interaction and intervention with corrections to the operation of the machine.

This mutual control is achieved through *communication*: each component controls the other by exchanging information with it and then in-forming it literally. The flow of information describes very well the process of communication between machines, but the same happens when a machine component and a human component are involved; the difference is that in the latter case the nature of the inputs is different and therefore *translation* is required. For this reason, it is not enough to understand agency in terms of intra-action: different parts of a system speak different languages, need interfaces to interact and communicate with each other, so that intra-action arises in the system.

The mechanism that concretely realises control through communication and that in fact turns bidirectionality into

[25] W. Ernst, *Technológos in Being: Radical Media Archaeology and the Computational Machine*, Bloomsbury, New York-London-Oxford-New Delhi-Sidney 2021, p. 131.

circularity (and thus creates a new level of causality that is unthinkable at the level of individual components) is feedback. Feedback causes a component to change itself according to a particular response it has received in an exchange of information with one or more other components. The process takes place in a circular exchange and leads to self-modifications based on the flow of information rather than direct modifications where one component is passive and another is active and performing. If we look at single "pieces of apparatus," elements within a technical individual or simple interactions where there is only one controller and one controlled, then we still observe linear causality[26]. However, if we turn our gaze to the complex system or observe points of intersection and exchange such as interfaces, we can recognise that the regime underpinning the entire interaction is that of *circular causality*. This not only allows us to rethink the role of the individual components of the system, but it is also what makes the system function more efficiently, because the feedback serves to reduce the system's dependence on the properties of its components[27].

6. *A new conception of determinism*

The recovery of the original theoretical project of cybernetics helps us retrieve the concept of determinism, confirming both that effects have causes and can *retroact* on causes, leading to new effects. This explains the apparent autonomy of AI systems as an emergent property of technical objects. Determinism, in light of information and circular causality, is evident in both biological and artificial systems, including AI.

This form of determinism allows for contingency and unpredictability at the system causality level, while saving the predictability of the effects at the linear and simple interaction

[26] See N. Wiener, *op. cit.*, pp. 97-98.
[27] *Ibid.*, p. 108.

level. Individuals, whether human or machine, are controlled by mechanistic processes but are *autonomous* to varying degrees. Autonomy, however, should not be equated with will or intentionality: it opposes automatism. The latter is pure mechanism, pre-programmed and hetero-directed. Many machines are automatic, automatic processes even take place in AI systems, and humans, animals or plants carry out many tasks themselves fully automatically, sometimes even after they have been trained to do so by other components of the system. Autonomy, on the other hand, requires feedback: an autonomous individual is able to turn its behaviour into an object that can be acted upon by modifying it according to previously achieved or unachieved results. For this reason, I support Yuk Hui's proposal to replace the term "autonomous" with the term "reflexive."[28]

Gotthard Günther described the cybernetic process of reflexivity as a "third transcendence" between pure subjectivity and pure objectivity[29]. It is an encounter that can never be realised between subject and object, but becomes an awareness of the gap between them and thus enables feedback between these poles[30], a concrete realisation of the Hegelian reflective logic[31].

---

[28] Y. Hui, *Introduction: Philosophy after Automation?*, in «Philosophy Today», LIV, 2, 2021, pp. 217-233, p. 218. Hui joins Simondon here, who describes automation as the lowest level of technical perfection and instead uses the cybernetic concept of reflexivity to speak of the highest level of technical perfection (see G. Simondon, *Technical Mentality*, in «Parrhesia», 7, 2009, https://www.parrhesiajournal.org/parrhesia07/parrhesia07_simondon2.pdf, last accessed: 15/01/2024). Hui also notes that the practical legacy of cybernetics may have led increasingly in the direction of an attempt to automate reflexive processes, and thus an overlap between autonomy and automation; but Hui, like me in this paper, seeks to recover the theoretical legacy of cybernetics, rather than its later applications.

[29] G. Günther, *Das Bewußtsein der Maschinen. Eine Metaphysik der Kybernetik*, Agis-Verlag, Baden-Baden 1957, pp. 30-32.

[30] Id., *Can Mechanical Brains Have Consciousness?*, in «Startling Stories», XXIX, 1, 1953, pp. 110-116. For Günther, the process of reflection is identified with consciousness, a problem that my paper tries to avoid by implying that it is misplaced and that it is worth re-reading these questions instead in terms of agency. It should be noted, however, how Günther has read the problem of consciousness in eminently operational terms: memory retains (unconsciously) the original impression ("a rose"); it adds (unconsciously) the I and a perception ("I see a rose"), but this is not yet the conscious state about a rose; the brain then compares the first message ("a rose") with the

Machines in Günther's time had not yet reached a perfect level of reflexivity, but AI systems today, like AlphaZero, have achieved a high level of reflexivity through unsupervised learning.

When it comes to reflexive machines, it is useful to ask whether one can speak of an *inner purpose*. Inner purpose does not mean entelechy in the Aristotelian or post-Aristotelian sense[32]. Every individual carries within it a pre-individual potential that still needs to be structured[33]. There are different inner purposes – or entelechies, always in plural form[34] – depending on whether the component is human or mechanical. The inner purpose of an AI system is algorithmic thinking, which involves tackling problems and solving them in a finite number of steps, applying strategies and recalibrating these strategies based on the results of actions.

It is not the algorithm *per se* that gives reflexivity to the machine: the algorithm «only expresses abstract thinking, and gains a quasi-autonomy when it is realized in machines»[35]. It is therefore the encounter of the algorithmic form of thinking with the particular architecture of digital technologies or artificial

---

second ("I see a rose") and discovers a non-equivalence, a surplus (the "third transcendence"). According to the author, this type of consciousness would be technically reproducible – self-consciousness, on the other hand, would not. In a way, this operational reconfiguration of consciousness could also be read as a reduction of consciousness to agency.

[31] See Y. Hui, *op. cit.*, p. 231.

[32] Entelechy (ἐντελέχεια) means the tension of an entity towards its perfect realisation according to its own laws and therefore inherent (see Aristotle, *De Anima*, Focus Publishing, Newburyport 2011, pp. 48-49). The term is sometimes understood as synonymous with actuality (ἐνέργεια), although the latter seems rather to denote the process of actualisation of a form, whereas entelechy would denote the perfect realisation of a substance already inscribed in the entity in the form of possibility. The same term is used by Leibniz to describe the monad (see L. Strickland, *Leibniz's Monadology. A New Translation and Guide*. Edinburgh University Press, Edinburgh 2014, p. 27), since it carries within itself the perfect organic purpose of its development. In any use, however, this term presupposes the idea of one and only one final cause, of a linear development towards a single possible, already predetermined purpose. Such an idea would contradict the form of determinism in the light of the notion of information, which is explained in these pages.

[33] See G. Simondon, *Individuation…*, cit., p. 183.

[34] See *ibid.*, p. 258.

[35] Y. Hui, *Algorithmic catastrophe–the revenge of contingency*, in «Parrhesia», 23, 2015, pp. 122-143, p. 131.

neural networks that gives AI autonomy or reflexivity: «Algorithmic thought itself has become an agency within architectural design processes»[36].

The operational complexity of the algorithm «overwhelms the simplicity and clarity of algorithmic thinking»[37], generates unforeseen and unpredictable effects, makes artificial agents autonomous, but connected to each other, to the environment or to other agents through the feedback mechanism. This agency, autonomous and capable of determining courses of action, is the agency of individuals at the level of interaction. At the level of intra-action, on the other hand, it can be read as a "collective technical agency"[38] that expresses the realisation of the system's entelechies in a *transductive process*[39].

Although the transductive process preserves a tendency given by entelechies (and the basic tendency is to maintain the circulation of information for the survival of a system), the outcomes remain open because they depend on the inputs and the selection between these inputs according to the tendency. This is what Mead essentially calls «a natural teleology, in harmony with a mechanical statement»[40].

---

[36] W. Ernst, *Technológos in Being…*, cit., p. 135.
[37] Y. Hui, *Algorithmic catastrophe…*, cit., p. 132.
[38] W. Ernst, *Technológos in Being…*, cit., p. 181.
[39] «By transduction» Simondon means «a physical, biological, mental, or social operation through which an activity propagates incrementally within a domain by basing this propagation on a structuration of the domain operated from one region to another» (G. Simondon, *Individuation…*, cit., p. 13). This structuring operation is not simply the transition from potentiality to actuality, for at the origin there is a *tension* – which is a problematic and pre-individual tension in itself and not a tension between a given matter and form – that is resolved in an inventive act: «The extreme terms attained by the transductive operation do not exist before this operation; its dynamism stems from the initial tension of the system of the heterogeneous being that phase-shifts and develops dimensions according to which it will be structured; it does not come from a tension between terms that will be attained and deposited at the extreme limits of transduction» (*ibid.*, p. 14).
[40] G.H. Mead, *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago 1972, p. 6.

7. A new theory of action

Key concepts such as system, information, modulators/interfaces, metastability, feedback, control, and communication, within a deterministic framework, help us describe agency in neutral terms, paving the way for a systemic theory of action. This "neutral" theory of action can be found in a revised version of Mead's social psychology. Namely, his social behaviourism attempts to explain both acts resulting from primary stimuli and immediate responses – as in unreflective (human and non-human) animal actions, but also in mechanical systems – as well as deferred acts[41], where processing time allows a reflexive selection between stimuli and a choice between possible responses – as in many actions of conscious agents or those performed by digital systems or AIs.

Mead's inclusion of the act's beginning (and thus the stimulus as well as the attitudes) in the act itself[42] allows us to view the conditioning of one component on another as an appropriation of stimuli from the environment or from interactions. Although Mead accepts the idea of an individual mind rooted in the central nervous system[43], he sees it as a product of social interactions[44], enabling the theory of action to rely on the transmission of informative meanings through collectives.

The collective makes communication possible, understood as the transmission of meanings through information. «The existence of the collective – writes Simondon – is necessary for information to be significative»[45]. This is because receiving information means undertaking an individuation, that is creating «the collective rapport with the being from which the signal arises»[46]. This links communication and information structuring to action, particularly

[41] *Ibid.*, pp. 90 and ff.
[42] *Ibid.*, p. 5.
[43] *Ibid.*, pp. 98 and 116.
[44] *Ibid.*, p. 7.
[45] G. Simondon, *Individuation…*, cit., p. 344.
[46] *Ibid.*

to transindividual action, which «makes it such that individuals exist together as the elements of a system that contains potentials and metastability, expectation and tension, then the discovery of a structure and of a functional organization that integrate and resolve this problematic of incorporated immanence»[47].

Emphasising the collective allows us to consider not only individual unidirectional actions but also *inter*actions that generate new group behaviours (*intra*-actions), new forms of training and learning, and networks of relationships. If we extend the notion of the collective in a cybernetic direction, we can include media and machine components, mechanical environments, natural environments.

Agents, both human and non-human, operate in a context where they receive stimuli, process them, respond, and retroact on the source of the stimuli, generating interactions and reciprocal conditioning. All these actions are based on information exchange and must be contextualised within a system or collective.

In this theory, the main difference between human and non-human agents lies in the nature of the selection process. While the human mind chooses by using various mechanisms, a computer follows a strictly binary logic.

Drawing from Mead's theory of action, first-order cybernetics, and Simondon's philosophy, an action can be defined as purposeful behaviour aimed at solving compatibility problems and establishing metastable equilibria through information transmission. This definition applies to AI systems as well, which can successfully perform such actions, impacting the environment and human components of the system. Responsibility for these actions therefore cannot be fully attributed to the humans who conceived and designed the AI systems, but calls into question a level of responsibility that has to do with systemic intra-action.

[47] *Ibid.*, p. 339.

8. *Conclusion*

By combining insights from a constellation of different disciplines and thinkers, this paper has forged a novel perspective on the theory of action in complex systems. Key insights include the crucial role of feedback in mediating communication between human and non-human components, the dynamic interplay of information in shaping systemic behaviour, and the recognition that determinism can also involve unpredictability and openness at the system level. Furthermore, the notion of inner purpose, whether in humans or machines, proves to be a key concept as long as it is understood as underpinned by the transductive processes that govern individuation within a collective milieu.

This framework challenges traditional models by emphasising the central role of the collective in shaping meaningful communication and action. By extending the concept of the collective to different elements, including media, machines, environments, etc., it is possible to redefine action as fundamentally rooted in information and contextualised within a larger system.

Against this background, it is possible to think of autonomous or reflexive artificial agents. But does this mean that they are also directly *accountable* for their actions?

The fear of certain humanistic approaches to technology is that recognising artificial agency could lead to a reduction in human responsibility. The thesis that I would like to put forward in conclusion, however, is that human responsibility is not only not diminished by machine responsibility, but on the contrary is increased.

I am not just referring to the fact that there is a responsibility of the designer[48] or a responsibility of the end user. I am

[48] Goetze, for instance, states that computing professionals are ethically required to take responsibility for the systems they design, despite not being blameworthy for the harms these systems may cause (see T.S. Goetze, *Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement*, in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and*

referring to a concept of distributed responsibility[49]. It derives from the extended notion of distributed agency[50] outlined so far. But before explaining in what sense I am talking about an extension of the "human side" of this distributed responsibility, it is necessary to make a distinction between responsibility and accountability.

Responsibility typically refers to the operational and direct obligations that an individual or entity holds. It is a duty that can be shared among multiple parties. On the other hand, accountability is the obligation for outcomes, even when operations are delegated. It presupposes a foresight of consequences and, unlike responsibility, cannot be shared.

When Goetze[51] speaks of conditions of personal responsibility – which I would understand instead as conditions of accountability – he is referring to a *control condition* and an *epistemic condition*. The control condition states that an individual A must have been, in some sense, able to control whether event X occurred or not. The epistemic condition states that an individual A must know that X would have resulted (or could have resulted) from the actions that A has taken; alternatively, if A did not know that X was a potential result of their actions, it must be true that A should have known.

Neither of these conditions are met for the human individual who would be held accountable for the actions of an artificial agent.

Transparency (FAccT '22), Association for Computing Machinery, New York, pp. 390-400).

[49] A. Strasser, *Distributed responsibility in human–machine interactions*, in «AI and Ethics», 2, 2022, pp. 523-532 talks about distributed responsibility between humans and machines. Strasser, however, takes a gradualist view of moral responsibility in which the "full-fledged moral agent" is the one with consciousness, intentionality, and free will, whereas I have instead sought to link agency to much less anthropocentric conditions.

[50] Distributed agency is discussed in N. J. Enfield, P. Kockelman (a cura di), *Distributed Agency: Foundations of Human Interaction*, Oxford University Press, Oxford 2017. My notion is "extended" because, unlike that text, it also considers artificial entities and various material structures in the distribution of agency.

[51] T.S. Goetze, *op. cit.*, p. 392.

This is because "autonomous" artificial agents can exceed the human capacity for control and prediction.

The fact that our artefacts "surpass us" by acquiring agential autonomy should not surprise us, but rather warn us to be particularly careful: if I cannot predict all the effects of technological systems[52], I must still try to predict as much as possible, including catastrophe among design parameters (design for failure)[53], or I must try to keep a (cybernetic) space open for feedback on technological systems. This should be the moral imperative of technological design, especially with regard to the reflexive capacity of the artificial, and corresponds to a strive in extending human *responsibility* – while accepting that we cannot extend accountability.

If we think back to the SyRi algorithm case mentioned at the beginning, we cannot say from this perspective that the people involved in the development, implementation, and training of the algorithm are individually guilty of the false accusations. However, it must be recognised that there is a shared responsibility that encompasses all these individuals and the public power that should have controlled them. The relative agential autonomy of AAs, where the concrete possibility of effective interaction with the human component was not foreseen, generated intra-active dynamics that were detrimental to the latter. This case illustrates very well the need to extend responsibility in order to avoid injustice.

To summarise, in charting the course for future research, it is crucial to establish ethical design frameworks that are not merely guidelines but integral components of AI development. These

[52] This is what Günther Anders defines as "Promethean gap" (*das prometheische Gefälle*): «the ever-increasing *asynchronisation between humanity and the world of its products*» (G. Anders, *Die Antiquiertheit des Menschen I. Über die Seele im Zeitalter der zweiten industriellen Revolution*, Beck, München 1961, p. 16, my translation), i.e., the gap between the maximum of what we can produce and the maximum of what we can imagine, the latter being "shamefully small" in comparison to the former.

[53] See Y. Hui, *Algorithmic catastrophe…*, cit., pp. 131-132.

frameworks should guide technologists and policy makers alike, based on *technical culture* as a discovery of the isodynamic between human thought, action, and technicality[54], and on the nurturing of imagination as *moral fantasy*[55] to anticipate the far-reaching consequences of AI and proactively engage with its ethical dimensions. While recognising the autonomy of AAs, we must still extend human responsibility by fostering a culture of responsibility that transcends individual and collective boundaries, providing vigilant oversight, continuous assessment, and a willingness to course-correct.

FRANCESCO STRIANO is Researcher in Moral Philosophy at the University of Turin

francesco.striano@unito.it

[54] See G. Simondon, *On the Mode of Existence…*, cit., pp. 134 and 159 and ff.
[55] See G. Anders, *Die Antiquiertheit…*, cit., pp. 271 and ff.