© Lo Sguardo - rivista di filosofia
N. 34, 2022 (I) - Algoritmo

DOI: 10.5281/zenodo.7770053

Articoli/2

Algorithm and Learning

Ethical and Theoretical Hypothesis on Empathic Robotics and Social Robotics

Igor Pelgreffi 0000-0002-4490-009X

Articolo sottoposto a doppia blind peer review. Inviato il 31/08/2022. Accettato 17/09/2022.

The article explores the relationship between algorithms, machine learning and artificial empathy. The pivotal problem is whether and how artificial learning of emotions by artificial agents, guided by algorithms, is possible. The first question consists in an in-depth analysis of the concept of learning (through some references to anthropotechnics, to Aristotle, to Merleau-Ponty and to Bateson), showing that – even in the human field – learning involves the use of forms of automatic and procedural repetition, such as the body schema or the behavioural algorithm. The hybrid aspect (between technique and nature) of learning processes in a broad sense is therefore underlined. The second issue is emotional learning by artificial agents. Through the use of an extensive specific bibliography, especially in the IT-engineering field, the possibilities and limits of this possibility are shown, discussing the major knot, namely that the machine or the algorithm has no living body (*Leib*), but mechanical (*Körper*). In the artificial field, the repetition and learning of patterns appear to be very far from that of human beings. However, the article tries to stress the usual categories and to formulate hypotheses, highlighting the possibilities of heterogeneous hybridization between the repetition learned in the machine and its ability to complexify its own behaviours, making them much less dissimilar from those of natural agents. The third question is that of the different ways in which the 'affective' dimension of the machine can be interpreted, through a relational systemic approach (social robotics) in which artificial empathy takes shape only within the relationship with the other. Different topics are explored, from those of emotional grafting (emotional chip) to that of a form of autopoiesis in the direction of a form of artificial autonomy, today in a larval but not absent form, through references especially to Dumouchel and Guattari. In the concluding part, some hypotheses are formulated on a conceptual frame for managing the ethical issue of autonomy, potentially also on an emotional level, of artificial agents guided by algorithms, suggesting an attitude capable of thinking of an ethics of hybridization but at the same time a hybridization of the usual ethical attitudes, that goes beyond the opposition between roboethics and machine ethics.

***

1. Algorithm and Learning: a background

The aim of this paper is to analyse the relationship between algorithm and automatic learning in the field of artificial empathy, but also to explore a number of hypotheses, that expect to be verified with further studies. Algorithm will be meant as a device that processes information in an increasingly organic and holistic way, rather than linear or serial, as in the field of *deep neural learning*, therefore as a device capable of solving *totally new* problems by learning from the environment, operating with increasing complexity, up to the extreme level: algorithm *that controls parts of the algorithm itself*. We reach here, therefore, the topic of *operational self-reflexivity* of the calculation machine, if not that of *machinic autonomy*. This last represents an extreme level of the questions. It concerns the *vexata quaestio* of the distinction between functioning and behaviour. After all, this is the philosophically central knot: will artificial agents be able to emancipate themselves completely? Will programmed algorithmic systems become autonomous through self-learning processes? The theme of *artificial autonomy*, recently re-proposed by Luca Fabbris[1], seems to be the only approach that leads to a real thought of AI understood as self-generation both of its behavioural patterns then of *autopoiesis* of its own ends[2].

But a further question should be introduced: can artificial systems learn an emotion? Can they gain *empathic autonomy*? The question, unlike that of autonomy in the computational field, involves the question of the body: can artificial agents feel emotions, if they have no living body (*Leib*)? The answer seems to be negative. If 'thought' for the machine today seems the undecidable, 'having a body' seems the impossible. Nonetheless, this article will suggest some investigative hypotheses on empathic learning by working on the 'margins' of the body-machine relationship, that is on borderline issues, still not fully codified and not even fully clear in their very definition.

Two preliminary questions. The first. Although the decision-making skill [3] of an algorithm still remain in the field of predictability or preventive programming, albeit stochastic, they collide with an *uncanny* empirical reality: programmers are unable to fully explain the behaviour of the algorithm, which becomes similar to a *black boxes model*[4]. Proof of this is the fact that a specific field of

[1] L. Fabbris, *Il programma, la rete, il circolo: un approccio Io-fi all'Autonomia artificiale*, «aut aut», 392, 2021, pp. 63-76.
[2] About a definition of autonomy, cf. also W. F. Lawless et al (eds.), *Autonomy and Artificial Intelligence. A Threat or a Savior?*, Cham 2017; J. Liu et al (eds.), *Autonomy Oriented Computing. From Problem Solving to Complex Systems Modelling*, Kluwer Academic Publisher, New York 2005; M. Prokopenko (ed.), *Guided Self-Organization. Inception*, Heidelberg 2014.
[3] On the value of the 'algorithmic decision', cf. at least G. Tamburrini, *Etica delle macchine*, Roma 2000 and, with extensive bibliography, F. Fossa, V. Schiaffonati, G. Tamburrini, *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Roma 2021.
[4] E. Esposito, *Dall'intelligenza artificiale alla comunicazione artificiale*, «aut aut», 392, 2021, pp. 20-34.

research is born under the name of *Explainable IA*[5]. The shift (historical, material: empirical) from single algorithms to huge extended algorithmic systems, which are massive computational machines[6], has among its consequences the existence of outputs not entirely derivable from inputs. The loss of total intelligibility represents a drastic break in the classical-modern logic at the base of algorithmic programming: between artificer and artefacts a relationship begins to emerge that is no longer only univocal or linearly causal. All this evidently opens up to great problem in the ethical field[7], promptly addressed also by the international guidelines for the ethical programming of intelligent machines[8].

However, the heuristic emphasis entrusted to concepts such as transparency, explainability, explicability or understandability[9] seems today to be more of a 'defensive' than 'normative'[10] character with respect to a technological progression that confronts us with outcomes that are not entirely controllable. The understanding of an unexpected or divergent 'algorithmic event' occurs *ex-post*: it would therefore not be a real understanding, but a rational justification possible only after the 'decision' of the artificial agent has already taken place. According to Elena Esposito, in the case of «deep learning algorithms, [*this*] is a fundamental obstacle: if by explanation we mean a procedure that allows human observers to understand what the machine does and why, the undertaking is hopeless»[11].

The second question concerns the relationship between machine and emotion: can an algorithm learn emotions? If we accept a general conception of emotion as an «inner process aroused by an event-stimulus relevant to the

[5] Concerning *Explainable AI* concept, see S. Whatcher, B. Mittelstadt, L. Floridi, *Transaparent, Explainable and Accountable AI for Robotics*, «Science Robotics», 6, 2017; D. Weinberger, *Machines Now Have Knowledge We'll Never Understand*, «Wired», 18.4.2017; F. Doshi-Velez, B. Kim, *Towards a rigorous Science of Interpretable Machine Learning*, arXiv:1702.08608v2, 2017; T. Miller, Explanation in *Artificial Intelligence. Insights from the Social Sciences*, «Artificial Intelligence», 267, 2019, pp. 1-38.

[6] N. Bostrom, *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, Oxford 2014.

[7] S. Vallor, G. A. Bekey, *Artificial Intelligence and the Ethics of Self-Learning Robots*, in Lin et al, *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*, Oxford University Press, New York 2017, pp. 338-353.

[8] Such as: EU, *White Paper on Artificial Intelligence. A European approach to excellence and trust*, Bruxelles 2020; High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI*, EU Guidelines, Bruxelles 2019 OECD, *OECD Council Recommendation on Artificial Intelligence* (2021), OECD, Paris 2021 *e* OECD, *Principles on AI OECD*, Paris 2020.

[9] A. Ienca, E. Vayena, *The global Landscape of AI Ethics Guidelines*, «Nature Machine Intelligence», 1, 9, 2019, pp. 389-399.

[10] The fundamental study still remains M. Ryan, B. Stahl, B., *Artificial Intelligence Guidelines for Developers and Users*, «Journal of Information, Communication and Ethics in Society», 19, 1, 2021, pp. 61-86.

[11] E. Esposito, *Dall'intelligenza artificiale alla comunicazione artificiale*, cit., p. 28.

interests of the individual»[12], therefore a physical dynamic linked to changes in the body and its neuro-physiological state[13], the answer can only be negative.

What does the machine really learn through increasingly accelerated repetition cycles, in training and self-correction that is progressively effective and adaptive? It only learns to *register* stimuli, to *recognize* emotions, to *categorize* frames, even if they are complex or 'open': the machine has no body. This also explains why in the setting of studies on artificial empathy[14] the dichotomy between two types of emotions is often affirmed: internal emotions and external emotions, where the former would be emotions that arise internally to non-artificial agents (typically humans and animals) and that the machine records and decrypts; the second would be those reproduced mechanically by artificial agents, typically shown by them to human beings. The opposition would therefore be between «genuine human emotions and simulated robotic emotions»[15]. In the following we will show how one can, cautiously and on the basis of a series of empirical and laboratory findings, attempt to deconstruct the internal/external partition, if one follows a more interactive, open or 'organic' model of Learning in the empathic field and if one he is willing to take on more complex epistemological but also ethical models. Before reaching this point, however, let us dwell on some aspects of the concept of Learning.

2. Learning and repetition

2.1 Functioning or behaviour

Speaking about algorithm that learns, are we not anthropomorphizing, that is, assigning the ability of 'learning' to something that by its nature cannot learn, but rather only re-produce, copy or execute what has already been 'learned'? Are we not naively applying a human lexicon to an artificial agent without a living body (*Leib*) but with only a mechanical body (*Körper*)? And even if it is legitimate to speak of learning in a computational machine, what is the object of the process? Of course, the machine purchases new functions, new protocol

[12] See «Emozione», in «Enciclopedia Treccani on-line» [12.3.22].
[13] This article will not examine the different theories of emotion, nor the nuances around the emotion/feeling distinction. As a basic orientation, *feelings* will be understood as elements more displaced, compared to emotions, on a subjective level, which stabilize external emotional stimuli over time; *emotions* instead as mechanisms closest to the psycho-bio-chemical level of the organism's response to the environment. Emotions and feelings will be understood, that is, as distinct but not ontologically independent, according to a line present, among others, in Antonio Damasio. Cf. A. Damasio, *Looking for Spinoza. Joy, Sorrow and the Feeling Brain*, HMH, Montreal 2004.
[14] As a reference, much used in this article, take P. Dumouchel, L. Damiano, *Vivre avec les robots. Essai sur l'empatie artificielle*, Seuil, Paris 2016. See also D. A. Norman, *Emotional Machines* in Id., *Emotional design*, Basics Books, New York 2004 and S. Schmetkamp, *Understanding AI. Can and Should we Empathize with Robots?*, «Review of Philosophy and Psychology», 2020, 11, pp. 881-887.
[15] P. Dumouchel, *Vivre avec les robots*, cit., p. 26.

segments, new automatic reactions to environmental stimuli: new 'machine fragments', in the pragmatic sense of machinic operation. That said, can we really claim that the machine learns new behaviours? Here again, the problem of anthropomorphization looms: the term behaviour is widely used in scientific and technical-engineering literature, but on a philosophical level it denotes a problem. Perhaps it would be more correct to use the term *functioning*? But, even here, it is quite clear that the lack of a correct distinction between *functioning* and *behaviour* can make any ethical reasoning on Learning very unsteady, as its foundations are unstable. So let's try to go in the opposite direction: beyond the easy man/machine mirroring effect – typical in these topics – let's try to reflect on a technical, if not 'machinic', component *latu sensu* present in human learning.

Since ancient times, philosophical thought on the meaning of learning has tried to define or contain an over-determined experience. Learning as reminiscence, how to remember something already 'written' (programmed?) but forgotten, as for the ἀνάμνησις in the *Meno* of Plato? Or learning as a recording, in the 'memory of the body', of functions or behaviours, of motor-patterns halfway between natural and cultural, as the notion of *hexis* is sometimes read in Aristotle, for example in the *Nicomachean Ethics*? Let's take a leap in time: in contemporary theories of learning in Cognitive Sciences, a widespread concept understands learning as a modification and acquisition of knowledge or behaviour. On the philosophical level, the notable point of this conception is to think together, that is, on the same logical level, *modification* and *acquisition*. How is this possible? Generally speaking, if I acquire an object, I do not necessarily change the behavioural pattern. But it is also true that if I modify behaviour, it does not mean that it is also incorporating an object.

As a first approximation, then, the heart of the learning process lies precisely in the *hybridization* between modification and acquisition, between insertion (passive) and modification (active). Here too, we find (as in all definitions of learning) an over-determined element on the logical level, which however on the practical-concrete level synthesizes in itself its own form.

We cannot dwell on these aspects; however, whatever the definition assumed, in learning it cannot be excluded an important, if not decisive, reference to the figure of repetition. Hanging up means being able to manage a relationship with repetition. Repetition is present not only in the reification, but also in the process, that is, in the 'becoming other' of a scheme (cognitive, bodily, cultural): in every experience that leads to having learned the object, repetition is essential.

Learning processes shows a close relationship with that of reiteration. This is likewise observable by recovering some layers of our philosophical tradition, in particular on the themes of doubling, of the *habit* of performing an action, of *second nature* understood as deep incorporation of *procedural automatisms*. In fact, it has been a debated issue since Aristotle the repetition of behaviour or the repeated practice as the key to conceiving a structuring of the ethical character,

for example, which leads to being-social agents. In *Nicomachean Ethics*, the ethical value of habit (*hexis*) was that of a *repeated action* can lead to virtue (*arete*)[16]. The *hexis* is an *embodied* and *embedded* disposition which also has its own transformative power, for example of an action that is not immediately moral in an action that is structured and consolidated as ethics, in the practical domain and in the social use[17].

The achievement of automaticity in the repeated gesture means that the learned gesture 'becomes natural', embodied as a second nature: and all this already means speaking of automaticity, of activation of a passive form, that is of activation in the human agent, *mutatis mutandis*, of an algorithmic procedure.

2.2 Technique and nature

The boundary between nature and technique shows itself – in the construction of the agent as an entity halfway between natural and artificial – much less rigid. To a certain extent, if the *hexis* can be said to be natural, it is because *it completes* human nature, extending the 'virtual-algorithmic' element with which the man is 'naturally' equipped. Learning then also becomes synonymous of experiencing, of living an active/passive dimension (also typical of a highly evolved artificial algorithm which, as mentioned, can even self-modify in the Learning process). But this transition can take place, from the point of view of the anthropotechnical question – from Mauss to Sloterdijk[18] – only thanks to the repeated exercise. From this point of view, the human being also learns to act *by applying a program*, exploring new input-output combinations, selecting some solutions over others and, in that, habit is a 'shaping' character – it *gives shape* – of our being. Shaping should be read in the sense of a future behavioural frame that self-codifies and stabilizes, orienting future choices. Somehow man needs *routine*, not to deny but to cohabit with his nature as an automaton, so that in the exemption of 'lower' mechanical functions, linked to procedural automatism, 'good' actions are performed without friction, automatically: in a natural way, as if a new physics of the body or a new form of the substitute were in play. The theme of the Aristotelian *hexis* therefore denotes the matter of procedural automatisms, which are nothing more than executable programs by now acquired but whose 'artificial nature' we have forgotten, and this is visible in tracery in that long path which leads to the heart of that reflection on the technique of the twentieth century: the topics of repetition and difference in French Though (Deleuze, Derrida), of alienation (Adorno, Horkheimer, Pollock,

[16] See Aristotle *Nicomachean Ethics*, tr. by R. Crisp, Cambridge University Press, Cambridge 2000, 1144a 26-28.
[17] *Ibid.*
[18] M. Mauss, *Techniques of the body* (1934), tr. eng. B. Brewster, «Economy and Society», 2, 1973, pp. 70-88 and P. Sloterdijk, *Du mußt dein Leben* ändern. Über Anthropotechnik, Frankfurt 2009.

Benjamin), of man as a *naturally artificial being* or even *naturally technical*, in German philosophical anthropology (Gehlen, Anders, Plessner).

However, repetition is also one of the features of every technical object: basically, technique *is* repetition. The graft in us of procedural automatisms is, indeed, the greatest sign of our 'technicalization'. The latter is, in turn, what allows us to reach skills and levels of expertise, to be ourselves 'good algorithms', that is 'effective' and is not something that is opposed to a supposed natural genuineness of the human being. In the wake of some typical findings of contemporary French epistemology, from Canguilhem to Simondon up to Derrida, we could say that between technique and body, between repetition and nature, between protocol and living body, no clear separation should be understood, but rather a sort of original reciprocal overlap: «The relationship between physis and technique is not an opposition; from the very beginning there has been *instrumentalisation*. The term 'instrument' is inappropriate in the context of original technicality. However, a repetition prosthetic strategy inhabits the very movement of life: life is a process of self-replacement, the passing away of life is a *mechanike*, a form of technique. Not only, then, is technology not in opposition to life, but it haunts it (*hante*) from its first beginning»[19].

2.3 Learning to unlearn

It must be noted that if learning deals with repetition, it deals also with a kind of repetition that is not trivially identical or merely mimetic: learning is indeed repeating, but in some way it is a repeating differently, a reiterating diverging in a way that, superficially, we can define as 'not mechanical'. Learning authentically also means *learning to unlearn* as well as *learning to resist*[20] to our own automatisms: the strength of the built-in algorithm. But what do we mean by the expression 'not mechanical'? Is there really something not mechanical in learning? Partly yes and partly no, as we have already shown limited to the theme of *hexis*, of repeated action in the anthropotechnical exercise, and gradually we could go as far as the theme of adaptive reinforcement or that of the formation of *body patterns*.

In learning, repetition is intrinsic. But is it an identical or mutable repetition? These are over-determined questions. Whether there is a non-repetitive element, we don't know. Generally, we are led to reply positively, and perhaps the problem lies in understanding the reason for this apotropaic reflex to deny the coexistence of inert repetition and living repetition, of identity and difference in learning. We should say that neuroscience[21] has not yet established

[19] J. Derrida, *Nietzsche and the machine*, «Journal of Nietzsche Studies», 7, 1994, pp. 7-66.
[20] From this point of view, some recent studies by Houdé are important, such as O. Houdé, *Apprendre à résister*, Le Pommier, Paris 2017.
[21] See J. Bernacer, J. A. Lombo, J. I. Murillo (eds.), *Habits: plasticity, learning and freedom*, «Frontiers in Human Neuroscience», 2015 (on line); I. Testa, F. Caruana (eds.), *Habits. Pragmatist Approaches from Cognitive Science, Neuroscience, and Social Theory*, Cambridge 2020.

certainty about all of this. Neuroeducation[22], in turn, reports solutions to this problem which, in the progress of its discoveries, gradually appear less and less certain or unambiguous. Phenomenology of the body, which have pointed, since Merleau-Ponty, to an idea of body learning intrinsically capable of producing repetition and difference, also does not completely solve the problem. Every organism, thanks to fundamentally repeatable physical patterns and structures, learns to diverge: learning «never consists in being made capable of repeating the same gesture, but of providing an adapted response to the situation by different means. Nor is the response acquired with regard to an individual situation. It is rather a question of a new attitude for resolving a series of problems of the same form»[23]. Note also: the greatest difference between the living organism and the machine would lie precisely in the operational and functional capabilities of the *Leib*, which while processing the input transforms it into 'other' than being the mere premise of an automatic reaction: it also diverges from its presumed datum identity: «the organism [...] cannot be compared to a keyboard on which the external stimuli would play and in which their proper form would be delineated for the simple reason that the organism contributes to the constitution of that form»[24].

All this because, in the phenomenological setting, we not only have but also are a body, a living body, a *Leib* that cannot be reduced to its *Körper* 'machinity'. But in reality, other readings of the phenomena of productive or generative learning are also possible that focus more on the fact that the living body is both mechanical (*Körper*) and not mechanical, that is, both mechanical and, at the same time, *more* than mechanical or *less* than mechanical. The *Körperschema* is the *sinolus* between scheme, mechanical protocol and adaptive plasticity to the environment, and we can therefore advance the hypothesis that the 'body-that-learns' is structurally not reducible, on a theoretical level, to one of the two regimes.

The body-that-learns is therefore operation and function, both vital and algorithmic. It is something qualitatively different from the expression of a simple mechanism, on the one hand, and from a field of 'freedom', of absolute release from physical-mechanical needs on the other. From this point of view, the body-that-learns is always an element of indeterminate mediation: an hybridization, a process halfway between Mechanism and Vitalism. The theoretical references could touch here on the issue of the intrinsically complex dimension of the repetition system in the living (but also in cybernetics) found in the works of Gregory Bateson in *Steps to an ecology of mind*. [25] The body-that-learns

[22] On neuro-education, see O. Houdé, *La neuroéducation: magie ou science? Cerveau & Psycho/ Pour la science*, «Chronique. L'école des cerveaux», 86, 2017, pp. 80-83 and F. Ramus, *Neuroéducation et neuropsychanalyse: du neuroenchantement aux neurofoutaises*, «Intellectica», 2018 1-2, 69, pp. 289-301.
[23] M. Merleau-Ponty, *The Structure of Behaviour* [1942], Beacon, Boston 1963, p. 124.
[24] Ivi, p. 13.
[25] See G. Bateson, *Steps to an ecology of mind*, Ballantine Books, New York 1971.

obviously acquires information, it 'feeds' on data; but it already does so by vitally interpreting those same data. But its real historical and material consistency lies in this ontological ambiguity. Bateson, as is well known, read cybernetics, the feedback and adaptive circuits of data processing systems (biological or artificial), in this very open perspective, in which ontological outcomes were never taken for granted, in a unitary theoretical framework that embraced in a consistent way the themes of learning, deutero-learning, unconscious learning and computational learning. In this line, it cannot then be excluded that even the elementary behaviour, or the pre-logical instinct, is in turn constructed or artificial, understood as a sequence of behavioural stages inherited in the body, and therefore basically an algorithm.

In short, learning means acquiring not only information, but 'plastic methods' for solving problems. The expression 'body-that-learns' is equivalent to that of 'data processor' to which, however, is added the idea of being also a powerful multiplier of incorporation of schemes and algorithms. Indeed, we can hypothesize that the body-that-learns, that is, the *technical operation* of *synthesis* that allows learning, is basically a form of hybridization between the scheme (even *Körperschema*) and the algorithm: one without the other, it makes no sense. And then it is worth remembering Simondon's words, on the value assumed by the concept of *technical operation*: «the technical operation is a pure operation that brings into play the authentic laws of natural reality; the artificial is aroused by the natural, not by the false or by the human taken for natural»[26].

Understood in this way, however, the 'open' and generative learning process abandons its anthropocentric quality, as even an artificial agent, within certain limits, can learn and participate in the *technical operation* of 'understanding the world'.

3. Emotions in the machine?

3.1 Emotion and learning

Can an artificial agent learn emotions? Let's start by seeing how the iterability component and the aspect of technical-nature hybridization seen so far, typical of human learning processes, are also very typical in the field of empathic and social robotics. And we remember how today, with different degrees, we are faced with robotic agents run by algorithms that learn to recognize, manage, reproduce stimuli attributable to emotions: sadness, joy, fear and their combinations[27]. Artificial empathy takes hold in very different fields: companion robots, care robots (elderly, temporary or permanent disabled people); robots for playing and for free time; robots to communicate in an increasingly effective and human-like manner; robots in the most varied activities: commercial, industrial, medical, sexual, military, educational. The set of these artificial agents interacts

[26] G. Simondon, *Sur la technique*, PUF, Paris 2014.
[27] For a broad and consistent review, see P. Dumouchel, *Vivre avec les robots*, cit., pp. 110-115.

with the human 'counterpart', simulating emotions, facial expressions, sounds, movements, postures increasingly similar to those of living beings, where repetition is crucial along the 'learning line' scheme, test, measure (or check), that becomes new scheme, new test, new measure, as confirmed in many studies, starting from those on the «affective computing» by Picard[28]. Again, something very similar to how humans also 'experience' the world.

Many ethical questions arise from the analysis of the historical-technological framework just mentioned. That a machine can, for example, imitate human emotions or recognize human emotions, raises some questions: why does it do this? For the benefit of whom (of the Alzheimer's patient who finds benefit from the stimulating and 'convincing' companionship of the empathic robot; or the robot factory itself; or the exemption from commitment by the patient's own family micro-network of social relations?). And again: how does it do it? What dangers and what advantages[29] does this technology of artificialization of emotion involve?

From a philosophical point of view, this kind of learning is about *representation*: a machinic apparatus recognizes a living thing by reducing it to an object, to a code, to memory. We are therefore still mainly on the representational or semiotic level – and certainly not corporeal or emotional – of learning. Here then returns the question, already foreseen, of the distinction between *internal* and *external* emotions[30], the first considered authentic, 'true', expressive of a 'subjective' psycho-physical reality; the second considered constructed, artificial, externalized by the machine without any feeling, without concrete expressive corporeality, deceptive[31].

But things are actually more complex than this framework might suggest. Many of these *intelligent artificial systems* begin to *autonomously* distinguish emotions. For the moment, these are necessarily elementary learning, which develop in ways that today are techno-physiological, biochemical or even bio-mechanical, little more than an artificialization of the reflex arc mechanisms. In other words, 'feeling affection' corresponds to a sort of small autonomous (but still mechanical) perception, to resume the Leibnizian lexicon of *small perceptions* as a zero degree of a process of subjectivation of being.

Within these limits, however, we could say – often without a real explainability – robots today *are learning to feel emotions*. And repetition plays its own role in the emotional learning process. This last, in its *initial* or *basal*

[28] See R. W. Picard, *Affective Computing*, M.I.T. Media Laboratory Perceptual Computing Section, Technical Report n. 321 (Boston 1995); L. A. Camras, *Expressive development and basic emotions*, «Cognition and Emotion», vol. 6, n. 3 and 4, 1992; I. R. Murray, J. L. Arnott, *Toward the simulation of emotion in synthetic speech*, «J. Acoust. Soc. Am.», vol. 93, feb 1993, pp. 1097-1108.
[29] N. Bostrom, *Existential Risks. Analysing Human Extinction Scenarios*, «Journal of Evolution and Tecnhology», 9, 2002.
[30] See D. Parisi, *Internal robotics*, «Connections Science», 16,4, 2004, pp. 325-338; Id., *Future robots. Towards a robotics science of human beings*, John Benjamins, London 2014.
[31] See P. Dumouchel, *Vivre avec les robots*, cit., pp. 91-132.

sequences, is nothing more than a form of coding: a) tracing certain sensitive data (images, tactile sensations, sounds) acquired by sensors within predetermined frames: categorizing the data (sad, cheerful, reassuring, etc.); b) replicate that type of emotion through complex purely mechanical dynamics. Thus far we find *passivity* (a) and *activity* (b). But there is no static stage, as the algorithm works immediately in the next level, that is that system of feedback on itself that allows the machine to understand that it has 'hit' or not the object (the target of 'correct emotion') with a suitably small error[32].

Precisely in this last stage a form of self-reflexivity seems to emerge as a *recognition* of having acted correctly. In turn, reinforcement in a very short time implements other 'affective acknowledgments', other simulated and tested emotions, other 'emotional behaviours'. But there is always – integrated – an inverse feature, as a sort of *reversibility* in the relationship: the machine must return a certain level of emotion to the human, producing the 'correct' emotions. Incidentally, this empathogenic quality raises, starting from Mori's studies in the 1970s, the ethical problem of excess machine-human proximity known under the name of *Uncanny valley*: androids increasingly similar to humans, in appearance, in behaviour, in emotional expressions[33].

Again: do all this mean that algorithms *learn to feel emotions*? That, that is, by repeating and complexifying the stimulus/response links, they gain a sort of *emotional autonomy* that even allows them to *produce emotion*, in the sense of an *empathic autopoiesis*? If learning deals with a form of repetition, can this feature be applied to the embodiment domain, that is, a 'progressive corporealization' of emotion in the machine? Let's make it clear: this is not the case today. The answer seems negative, or limited to the field of science fiction or cinema, from *Blade Runner* or *Matrix*, to literature, the machine that 'comes alive', which becomes autonomous from the automatic, in terms of feelings, as well as emotions. As in *Klara and the sun*, by Ishiguro[34], a totally 'subjective' novel: the robot Klara tells his story while it learns to feel emotions, to deceive, to suffer, to experience a true friendship with a human and, even, to desire.

3.2 Emotional chip

Shouldn't we ask ourselves why the answer seems negative to us? As to why, as already written in the introduction, if the 'thought' for the machine today seems an undecidable, 'having a body' seems the impossible (*res cogitans* versus an 'alive' *res extensa*)?

Let's start again. How can a machine feel emotions? Pausing a little longer in the field of Fiction, we cite the example of the *emotional chip* taken from the famed Star Trek series. There we find on the one hand Spock (human mother

[32] L. Pitt, L. Valiant, *Computational limitations on learning from examples*, «Journal of the ACM», 35, 1988, pp. 965-984.
[33] See M. Mori, *Bukimi no tami (The uncanny valley)*, «Energy», 7, 4, 1970, pp. 33-35.
[34] K. Ishiguro, *Klara and the sun*, Random House Large Print, London 2021.

and Vulcan father), who is cold and particularly clever, and on the other, Captain Kirk, who is often emotionally involved in choices and decisions: the crystalline reason of the former is dialectically opposed to the 'emotional reason' of the second. Now, the aforementioned dialectic between pure intelligence and bodily reason finds a sort of update in the following series, in the figure of Deputy Commander Data. Data is, in fact, a totally artificial android. The problem is then that of endowing it with an emotional component. To equip him with emotions, a chip is inserted into the circuits and therefore the algorithms that govern his behaviour.

In short, the idea is that emotion is, in turn, something programmable, 'algorthimizable' by means of a physical graft. Something that, however, seems to naively re-propose the mind-body split[35]. The example of the emotional chip, while being part of Science Fiction, is able to represent the basic idea underlying a whole series of empathogenic devices potentially developable in the future, that is an algorithm capable to read emotions but above all to give emotions: receptive passivity and productive activity, where the learned skill consists in incorporating a frame of emotions on an increasingly 'machinic' level.

But the fundamental problem of the emotional chip, as well as of all devices related to it, is that of its 'coming from outside', perfectly in line with the conceptual paradigm of an external emotion that would be ontologically and genealogically completely different from an internal emotion. It is a graft, where the artificial remains artificial, therefore manageable, programmable and controllable by man. In its being an external object, the chip confirms – as in a game of mirrors – that the machine remains a machine: only an 'other' element can animate the machine and give life: this quasi-theological aspect recalls other well-known scenarios, as in the dramatic dialogue in the *Blade Runner* film between the replicant Roy Batty and his human programmer, Tyrrel. The problem is that of *giving life* on the part of the creator: giving life to 'something' that obviously has no life and that is destined to remain forever only 'mechanical' but who, after realizing that his operational capacity is over, that is, that he has been programmed from the outset with a countdown that leads him to death, asks dramatically: «I want more life, father!»[36]. Is it a question of the human or of the replicant?

3.3 The hypothesis of artificial emotional autonomy

We then begin to ask ourselves whether a less 'dualistic' vision is possible or not in the theoretical field. Again: can the machine gain some kind of *emotional artificial autonomy* if it is, by definition, lacking of a body? We must be very careful here if we want to avoid falling back into the *doxa*, albeit millenary,

[35] However, note that there is a discussion on these aspects, as in R. Sekuper, R. Blake, *Star Trek on the Brain. Alien Minds, Human Minds*, Freeman, New York 1998 and D. A. Norman, *Emotional Design*, cit., p. 92 e ss.
[36] From the film: *Blade Runner*, by R. Scott, USA 1982.

which hypostatizes the existence of an ontological difference between machine and nature. The problem is now of a historical-material nature: it is 'between us', in its *uncanney* novelty and urgency.

Reversibly, even on the statute of the *Leib* some questions become possible today, in the line already mentioned in the analysis of the phenomena of *learning with repetition*: between mechanical learning and 'alive' learning, what exactly are the differences, if they really exist? In this sense, come empirical-experimental findings concerning the discussion of the dualistic algorithm-body separation deserve to be mentioned.

Part of them derives from an epistemological instance, that is, the need for a complexification of the cognitive aspects of robots in the context of the experimental development of *embodied* approaches. In this line, certainly as its extreme branch, the idea of recognizing robotic agents as *having a body*[37] has also made its way. This vision is obtained, precisely, through cross-disciplinary approaches that range from the perspective of *synthetic biology* to the concepts of *Organismically-inspired robotics*[38]. From this perspective, robots are seen as «complete agents»[39] as they are «endowed with a body» and, as such, considered capable of autonomously modelling a set of dynamic interactions between the body dimension and the environmental or systemic dimension[40]. This behavioural self-sufficiency, for now limited to specific fields, nevertheless represents a form of *primitive autonomy*. The decisive point of this research is the increasingly frequent adoption of 'systemic' and 'holistic' paradigms, an adoption that facilitates the epistemology of artificial empathy. The framework is that of very refined experiments, that of small differential iterations and minimal procedural integrations which, overall, implement and improve the 'sentient' algorithm. In the holistic-relational paradigm, *emotion in the machine* is therefore conceived as an emergency but always and only within a systemic dynamic, often resorting to the old ideas of Varela and Maturana or Bateson.

[37] On the question of the body in a robot, there are several studies, starting from the works of Ziemke, like T. Ziemke, *What's that thing called embodiement?*, in *Processing of the 25th Annual Meeting of the Cognitive Science Society*, Lawerence Erlbaum, pp. 1305-1310, 2003; Id., *The body of knowledge*, in L. Damiano, Y. Kuruma, P. Stano (eds.), *What can shyntetic biology offer to artificial intelligence (and viceversa)*?, «Byosystems», 148, 2016, pp. 4-11; Id. e R. Lowe, *On the role of emotion in embodied cognitive architectures: from organisms to robots*, in «Cognitive Computation», 1, 1, 2009, pp. 104-107.

[38] See E. Di Paolo, *Organismically-insipred robotics. Homeostatic adaptation and teleology beyond the closed sensorimotor loop*, in K. Murase, T. Akasura (eds.), *Dynamic Systems Approach for Embodiment and Sociality*, «Advanced Knowledge International», Adelaide 2003, pp. 19-42; T. Froese, *On the role of AI in the ongoing paradigm shift within the cognitive sciences*, in M. Lungarella et al (eds.), *50 Years of Artificial Intelligence, Lecture Notes in Computer Science*, vol. 4850, Springer, Berlin 2007, pp. 63-75; L. Damiano, P. Stano, *Synthetic biology and artificial intelligence. Grounding a cross-disciplinary approach to the synthetic exploration of (embodied) cognition*, «Complex Systems», 27, 3, 2018, pp. 199-228.

[39] P. Dumouchel, *Vivre avec les robots*, cit., p. 118.

[40R.] Pfeifer, C. Scheier, *Understanding Intelligence*, MIT Press, Cambridge MA 1999; R. Pfeifer, J. C. Bongard, *How the body shapes the way of we think*, MITPress, Cambdridge 2006.

In other words: the continuous self-corrections of the algorithm's course, the integration of the infinitesimal automatic and quasi-organic adjustments, make sense only if they are designed in constant coordination with the environment. There is therefore a need for an *interactionist approach*[41] through which to build artificial systems capable of *synthesizing* the two dimensions of emotions (internal and external). From this point of view, indeed, strictly speaking, there are no two types of emotions; what pre-exists is an *affective circuit*[42]. In other words, internal and external emotions are both necessary and derived from the construction of affective circuits, with the consequence that the distinction between natural systems for regulating emotions and algorithmic systems is becoming less and less definite.

Therefore, the artificial agent can 'learn to feel' and 'to be empathic' only if it is thought and designed within an intrasystemic dimension, autopoietic in the sense of the notion of *enaction*, that is, that form of activity/passivity as a figure of the agent within complex systems. The epistemological models of the 'radically incorporated' artificial mind, conceiving it as an *emerging entity*, therefore tend to overcome the dualism between an immaterial substance and an extended substance. Within this emergent paradigm, the artificial quality of emotion is determined as an emerging form from a «process of co-specification that dynamically connects the agent's nervous system, his body, environment and other agents»[43]. This dynamic is therefore the real protagonist of the automatic learning scenarios: it escapes the spatial logic of the separation between internal and external, laying the logical and epistemological basis for also overcoming the threshold between internal and external emotions, where the logical structure of *relationship* (between artificial and natural agents) is that of reciprocal intertwining or *embrication*, in the terms of Varela and Thompson[44].

Let us come back for a moment, however, to the theoretical value of the autonomy of an artificial agent, also recalling other methodological approaches. Rather widespread, it is the one that put into parallel the empathic learning by artificial agents with computational learning of cognitive procedures in the strict sense. In these parallels, another key issue often emerges, namely that of autonomy in its relationship with automatism. What, for example, would the learning algorithms characterizing AI lack today? Not the ability to calculate, nor speed or precision, but the ability to *give oneself a purpose* for one's own operations. In short, the question of the autotelic agent would be the heart of *artificial autonomy*, that is the real boundary to be reached and overcome in order to really talk about *algorithms that learn*. Once again, as was clear since Wiener's cybernetics, a parallel is imposed between the biological organism that learns and the program of the machine. Why is an organism not a machine,

[41] See P. Dumouchel, *Vivre avec les robots*, cit., p. 129.
[42] Ibid.
[43] Ivi, p. 141.
[44] E. Thompson, F. J. Varela, *Radical embodiement. Neural dynamics and conscoiusness*, «Trends in Cognitive Sciences», 5, 10, 2001, pp. 418-425.

Georges Canguilhem also wondered, in the era of the affirmation of DNA and cybernetics? For the reason that «the body has [...] a wider range of action than the machine. Compared to the machine, it has fewer purposes and more potential»[45]. The organism certainly acts by virtue of a program, of a 'biological and genetic algorithm'; but at the same time, it can also modify the program, it can question its own goal, deny it, translate it, postpone it. As Fabbris argues, we can then speak of authentic AI if and only if it is endowed with artificial autonomy (AA), meaning the autonomy of an algorithmic system capable of autopoiesis of its own ends: «a system is not only autonomous when it is able to reach, without any external intervention, a certain purpose, but also when it is able to give itself a purpose. Only this last capacity allows to distinguish an autonomous system from a purely automatic system»[46]. Fabbris therefore is, implicitly, along a line we have repeatedly evoked of a radical change in the point of view: «AI never existed, there have been only good imitations of descriptions of behaviours considered by us to be an expression of intelligence. AI will not exist until an AA is reached»[47]. Something similar could then be argued about the question of the learning of emotions, including the question of the body and the automatic system as an artificial organism, underlining the emergent, dynamic or generative aspect of this type of algorithmic artifice, if it is true that strictly speaking, in the field of *Guided Self-Organization* as well as of *Autonomy Oriented Computing*, «you cannot program an AA system […] but you can generate it by simulating the process through which it is identified»[48].

What derives from these examples drawn indifferently from the experimental (in vitro) and real (*in vivo*) domain? A confirmation that internal emotions and external emotions are no longer defined by excluding each other, but *one in the other*. The emotional aspect of artificial learning would not be understandable without a reference to this complex, organic-systemic framework, in which emotion shows itself as primarily a relational reality, and not just autochthonous and individual.

Artificial emotion is not a state, but an emergency, itself a continuous correction and adjustment of *everted* processes, open to relationship, interconnected to the environment with which the machinic agent interfaces (both the natural and the artificial environment). Artificial emotion is a function internal to a field of bodily signification: a field of forces, that is, of actions and feedbacks, of new meanings of the activity/passivity couple of an artificial agent that dialogues with the natural and gradually imitates it better and better, it

[45] G. Canguilhem, *Machine and Organism* [1952], in «Incorporations» (J. Crary, S. Kwinter eds.), Zone Books, New York 1992, pp. 45-69.
[46] L. Fabbris, *Il programma, la rete, il circolo*, cit., p. 64.
[47] Ivi, p. 76.
[48] Ivi, p. 65. On the complex definition of autonomy, cf. also W. F. Lawless et al (eds.), *Autonomy and Artificial Intelligence. A Threat or a Savior?*, Springer, Cham 2017; J. Liu et al (eds.), *Autonomy Oriented Computing. From Problem Solving to Complex Systems Modelling*, Kluwer Academic Publisher, New York 2005; M. Prokopenko (ed.), *Guided Self-Organization. Inception*, Springer, Heidelberg 2014.

integrates it (under the guise of schemes), pursuing a sort of *autonomy in the relationship*.

4. Traces of algorithmic sociality

4.1 Social robotics

To complete and better verify this vision, let's take a look at the 'social' side of artificial emotion, exploring the field of social robotics, a research field in which converge Human-Robot Interaction (HRI), Affective Robotics, Cognitive Robotics, Affective Developmental Robotics[49]. The original question of social robotics concerns how it is possible to insert, in our relationships and social lives, devices guided by algorithms, such as robots or other automatisms, in a plurality of contexts (assistance, surveillance, information, education, etc.). For our interests, the question is how to create social robots that have a relationship with humans that is 'credible' or 'likely', in terms of bodily and emotional reactions, guaranteeing the so-called 'social presence' of robots. The algorithm should be therefore 'human oriented', that is, capable of causing, in the perception of the bodies with which it interacts, sensations and 'affects' comparable with, if not indistinguishable from, those of human relations, in the sense of what could be called an *affective performance*[50].

In our approach, the reference to sociality is significant above all for its contribution to the *evolution* of artificial empathy: the progress in the ability of robots to decode and reformulate learned emotions, in fact, helps to produce (both theoretically and practically) artificial models of emotional processes that are increasingly close to the 'true'. But social robotics proposes again the ethical and philosophical problem of the emotional automaton in a different way. It is now clear that the programming of the machines themselves is conceived in order to implement self-learning and self-adaptation skills: we now speak of «creation of emotional regulation systems»[51] for robotic agents, therefore of a form of autonomy in establishing hierarchies, in making 'emotionally oriented' decisions, in making the algorithm more adherent to the synesthetic complexities of the environment in which it operates. At this point, the self-

[49] See K. Dautenhahn, *Methodology and themes of human-robot interaction. A growing research field*, «International Journal of Advanced Robotics», 2007, 4, 1, pp. 103-108; C. Breazeal, *Emotions and sociable humanoid robots*, «International Journal of Human-Computer Studies», 59, 1, 2003, pp. 119-155 and Id., *Toward sociable robots*, «Robotics and Autonomous Systems», 42, 3-4, 2003, pp. 167-175.
[50] A. Paiva (ed.), *Affective Interactions: Toward a new generation of computer interfaces*, Springer, Berlin 2000; T. Fong, I. Nourbakhsh, K. Dautenhahn, *A survey of socially interactive robots*, «Robotics and Autonomous Systems», 42, 3-4, 2003, pp. 146-166; H. Li *et al*, *Towards an effective design of social robots*, «Int. J. of Social Robotics», 2011, 3, 4, pp. 333-335. On the philosophical problems that such realities raise, for an overview see I. Pelgreffi, *Ambiente digitale, automatismi e corporeità*, in (F. Miano, L. Alici eds.) *L'etica nel futuro*. Orthotes, Napoli-Salerno 2020, pp. 389-399, and Id., *Bernard Stiegler e la critica della società automatica*, in B. Stiegler, *La società* automatica, Meltemi, Milano 2019, pp. 11-26.
[51] P. Dumouchel, *Vivre avec les robots*, cit., p. 117.

sufficient, autonomous and only *internal* character of a human body of emotion begins to lose its supremacy in the logical-epistemological model.

As a last consequence, we see then that social robots operate today through increasingly differential and differentiated strategies with the environment and with other agents, not only human, but potentially also robotic. On a level of radical empiricism, we can even glimpse the possibility of a robot-robot learning relationship, that is a relationship in the machinic field, considering that communication between machines takes place more and more often, directly or indirectly, via the internet. The relationship, that is, is more open and more indeterminate than those programmed with the machine *in vitro*, allowing for artificial systems that are much more similar to an *organism that learns*, rather than to a *simple automatic repetition machine*. On these future scenarios of unprecedented hybridizations, Guattari wrote:

> Leroi-Gourhan emphasised that the technical object was nothing outside of the technical ensemble to which it belonged. It is the same for sophisticated machines such as robots, which will soon be engendered by other robots. Human action remains adjacent to their gestation, waiting for the breakdown which will require its intervention: this residue of a direct act. […]. Curiously, in acquiring more and more life, machines demand in return more and more abstract human vitality: and this has occurred throughout their evolutionary development[52].

Guattari's point, also representative of many others, must be understood along a line of development, certainly in the 'irregular' manner that characterizes the author, of the French epistemological *coté* that ranges from Simondon, Lyotard, Derrida to Deleuze. It testifies to an urgency: it suggests not a trivial elimination of the human in the presence of the machine, but an *accomplishment of the human* that passes through a modification of the entire reasoning on the *learning abilities* of algorithms and robots, that is, on the 'sense' of this capacity in relation to the human being, as well as on their ways of self-reshaping their internet connections, their transmission and sharing protocols.

Algorithms are experimenting today, every second, with other ways of communicating: they 'invent' new languages, new codes that will not necessarily be shared with their 'creators'. From this point of view, the question of Machine Learning always refers, even when it is not explicitly mentioned, to the theme of evolution, to a sort of 'evolutionary process' of which Guattari speaks, which involves machine and human within a inextricable and over-determined association, a relationship that *becomes new* and non-programmable.

[52] F. Guattari, *Chaosmosis. An ethico-aesthetic paradigm* [1992], Indiana University Press, Indiana 1995, p. 36.

4.2 Affection as a mechanism? Heterogeneity and relationship

Among the theoretical sides of social robotics, we find the crucial possibility of conceiving the *affect* as a mechanical phenomenon. Does this mean that artificial sensory stimulation can lead to a proto-emotion? The most correct answer, for now, is: we cannot know. Certainly, for this to happen, a sort of 'social dimension' of the machine must be integrated into the dimension of the 'stand alone' sensor.

This means that the whole emotional field can be understood as a «continuous mechanism of inter-individual coordination»[53]. The cooperating side between different fields (between artificial and natural; between robotic body and human body) interprets emotion as an *empathic operational ability*. The emotion arises in the coordination-with, that is in the relationship-between, and it is no longer just the result of a movement of 'introversion' or of 'stand alone' computational re-elaboration, or of a doubling by the artificial agent. Giving praise to this particular sociability of the machine also means, in perspective, accepting that there are shares of autonomy in these relations between machines; it means also accept a concept of *machinic assemblage*, so to speak, which, as anticipated, brings the problem back to the great discussion on the emergence of forms, on saliences, on pregnancies, along a line that from the studies of René Thom and Jean Petitot leads to recent perspectives, such as those, which deepen and revive those of Deleuze and Guattari, by Alessandro Sarti and other philosophers and scientists[54].

Under the name of the sociability of algorithms, we are witnessing today a free game of recombinations of forms, the outcome of which is not predictable or programmable, being rather more similar to a *heterogeneous becoming of forms*: a continuous metamorphosis of the machinic form (which at the same time recalls the natural form). Some of the most advanced approaches in empathic social robotics postulate, in fact, a sort of *cognitive heterogeneity*[55], according to which there is a plurality of human minds, animal minds, but also robotic minds, which 'coexist' together. And this therefore characterizes the field of relationships in which the first trace of artificial emotion will eventually take shape: a field that is actually an intersection of heterogeneous fields but able to coordinate and communicate continuously. In this perspective, the difference between 'minds' is not ontological qualitative, but lies only in their being objects (or better cognitive systems) emerging from different environments, with different dynamics, with historically different relationships and interactions.

[53] P. Dumouchel, *Vivre avec les robots*, cit., p. 21.
[54] Line clearly visible since form F. Berardi, A. Sarti, *Run. Forma, vita, ricombinazione*, Mimesis, Milano-Udine 2008, and recently deepened and modified in A. Sarti, G. Citti. D. Piotrowski, *Differential Heterogenesis. Mutant forms, sensitive bodies*, Springer, Berlin 2022, research based on a dynamic morphology of the living in a post-structuralist approach. See also the old but fundamental study U. Fadini, *Principio metamorfosi. Verso un'antropologia dell'artificiale*, Milano 1999.
[55] P. Dumouchel, *Vivre avec les robots*, cit., p. 24.

In conclusion: the research on social robotics opens glimmers towards a different thought of the *affective-machinic dimension*. Among these, perhaps the most relevant is that of a reduction in the dichotomy between authentic human emotion and simulated robotic emotion[56]. And this reduction derives, let us note, from considerations that are not so much theoretical as empirical: social robotics explores the *interaction* between inside and outside, progressively deconstructing – in the practical level – the dichotomy between natural and artificial. Take, for example, some old research by Paul Dumouchel[57], over time and also recently resumed and articulated with Laura Damiano[58], which effectively describe the picture we are sketching:

> it is a relational approach to the theoretical characterization of emotions that it identifies in the affective exchange based on emotional expression a mechanism of inter-individual coordination and mutual determination of the inclinations of action. The central hypothesis is that what we commonly call 'emotions' are salient moments, or points of balance, in a continuous process of coordination. From this point of view, emotions must be thought of as common works, resulting from interactions, rather than as internal and private events[59].

Starting from that, the whole question of the passive-mimetic aspect finds its completion in these theories towards a more active dimension: activity and passivity, which are naturally intertwined in the living organism, also begin in the robotic field to train each other in increasingly complex modes, which signify forms of 'relational autonomy'.

The thesis of *affective coordination*, if taken to the end, completely redefines the historical, logical and theoretical terms of the problem of empathic robotics, leaving us to imagine that there may be a zone of *independence* in the emergence of emotions, on a socio-relational basis, in the robot, and therefore of *autonomy* in the algorithms that govern them. An area of which we still do not know much on the theoretical level, but with which we are resonating on the level of historical and practical knowledge.

[56] See R. Sparrow, L. Sparrow, *In the hand of machines? The future of aged care*, «Minds and Machines», 16 , 2, 2006, pp. 141-161; A. Sharkey, N. Sharkey, *The crying shame of robot nannies. An ethical appraisal*, «Interaction Studies», 11, 2, 2010, pp. 161-190 and Ids., *Granny and the robots. Ethical issues in robot care for elderly*, «Ethics and Information Technology», 14, 2, 2012, pp. 27-40; S. Tisseron, *Le jour où mon robot m'aimera. Vers l'empathie artificielle*, Albin Michel, Paris 2015.

[57] See P. Dumouchel, *Emotions. Essai sur le corps et sur le social*, «R. Ph. de la France et de l'Etranger», 187, (1), 1997.

[58] See L. Damiano, P. Dumouchel, *Artificial empaty, imitation and mimesis*, «Ars vivendi», 1, 1, 2011, pp. 18-31; L. Damiano, P. Dumouchel, H. Lehmann, *Towards human-robot affective coevolution. Overcoming oppositions in constructing emotions and empathy*, «International Journal of Social Robotics», 7, 1, 2015, pp. 7-18

[59] P. Dumouchel, *Vivre avec les robots*, cit., p. 27.

5. Beyond the dichotomy: an ethics of hybridization or a hybridization of ethics?

Going to the conclusions. On the one hand, we have seen that if it is true that the issue of learning, in the field of AI, is connected to a computational dimension, it is also true that something different happens when we turn to Learning applied to the empathic field. In fact, here the topic of the body is somehow central, in the sense of a (problematic) corporeity of the machine. But this *latu sensu* corporeal dimension acquires meaning only, as seen, in reference to the social and systemic quality of the algorithm: emotions can be learned by the algorithm if and only if the concept and experience of emotion artificial is understood on a social and holistic basis.

The hypothesis therefore, which concerns a possibility of reading the near future, is the following: only starting from this relational field we might speak consistently of *machinic learning of affects*, of an *affectivity in the machine* (both even at an embryonic stage of evolution) and other formulas we have encountered.

On the other hand, we had instead emphasized the fact that, reversibly, even in the human behaviour of learning there is an automatic, repetitive, machinic component in a broad sense. And we see also how the automaton 'in ourselves', is not only to be relegated to secondary, but rather to be recognized, accepted and developed, as a completion of our being in its most comprehensive sense.

From the above, it follows the need to explore the hypothesis of a hybridization dimension: in both sides of the learning process – in the technical-artificial domain *and* in the human-natural domain – we find isomorphism and reversibility, at least on the level of epistemological models. They are two 'opposite' sides of a process that is one, in itself.

However, the great problem in the ethical field remains the following: how to make these concerns on artificial empathic learning adhere to the usual ethical frames, that is, those currently in the field in the main international guidelines on machine learning?

A possibility is to hypothesize a different ethical attitude, which starts from a slow, patient and shared attempt to deconstruct the dichotomy between natural and artificial, for example between internal emotion and external or 'machinic' emotion. As Damiano and Dumouchel suggest in their essay on artificial empathy, the assumption of a less 'dualistic' perspective around the theme of emotions, «transforms the configuration of ethical issues raised by the relationships between substitutes and human beings»[60]. There is, moreover, a historical datum, namely the emergence with increasing experimental evidence that «when artificial emotions and empathy are considered part of a coordination mechanism, they become something very different from deceptive surrogates.

[60] Ivi, p. 28.

They become the occasion for a possible *co-evolution* of new social partners, capable of being integrated into the fabric of our relations»[61].

It is well known that since the first international conferences, such as that of Sanremo in 2004 (and the related *Euron Roboethic Roadmap*[62]), two lines have emerged around the problems of learning and autonomy[63]. One is concerned with managing automation from the point of view of man: the ethics of machines is an ethics that clarifies the responsibilities of the man or the 'creator'. The other line, minority and more radical, is instead willing to give space to the autonomy of the artificial agent: the machine can become a 'subject', a subject of law but also an ethical subject, with its responsibilities and rights[64]. On the one hand, the ethics of machines therefore remains the ethics of the engineers, programmer or designer, separating the human from his 'object'; an object, if anything, to be ethicized (and where therefore the instance of management and control, the instance also of responsibility, remains firmly in the hands of the human). On the other hand, one can imagine an ethics where it is accepted – in a dystopian or utopian way – that the machine can acquire, by *learning from itself in view of itself*, a total autonomy: an *autotelia of artificial operation*. Even in this second case, however, we find, inside the logical structure of the problem, a separation between the human being and the machine: one wishes to ethicize the object; the other to de-humanize ethics.

The concerns collected in this essay, which are nothing more than a preliminary reasoning, cautiously suggest the option of a third model of ethics: a sort of ethics of hybridization, but also a sort of hybridization of the aforementioned ethical attitudes. There are already traces of it, if we think for a moment of the great tradition of philosophical anthropology as an anthropology of technology and of its basic assumption that sees man as a *naturally technical being*, or of the line of reflection that belongs to Donna Haraway, since from his 1985 *A cyborg manifesto*, where the challenge was to understand that non-humans exist not only in the field of animals or plants, as a form of biological life, but also in the sphere of the 'animated' inorganic, capable – within the

[61] Ibid.
[62] G. Veruggio, *Euron Roboethic Roadmap*, Release 1.2, january 2007 (www.roboethics.org).
[63] The specific bibliography on the 'dichotomy' is very extensive. See P. Lin, R. Jenkins, K. Abney, *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*, Oxford University Press, New York 2017; S. G. Tzsafestas, *Roboethics. A Navigating Overview,* Springer, Berlin 2016; Vallor, S. and Bekey G. A., 2017, *Artificial Intelligence and the Ethics of Self-Learning Robots*, cit.. G. Veruggio, F. Operto, G. Bekey, *Roboethics. Social and Ethical Implications*, in Siciliano, Khatib, *Springer Handbook of Robotics*, Springer, Berlin[2] 2017, pp. 2135-60; M. Anderson, S. Anderson (eds.), *Machine Ethics*, Cambridge University Press, Cambridge 2011. N. Bostrom, E. Yudkowsky, *The Ethics of Artificial Intelligence*, in *The Cambridge Handbook of Artificial Intelligence*, K. Frankish, W. M. Ramsey (eds.), Cambridge University Press, Cambridge 2014, pp. 316-334.
[64] A vision that, we should remeber, often encounters difficulties among scholars of ethical questions. See W. Wallach, C. Allen, *Moral Machines. Teaching Robots Right from Wrong*, Oxford University Press, Oxford 2009, or even the sceptical considerations of Luciano Floridi, for example those summarized in L. Floridi, *On algorithms: ethical and epistemological questions. Interview with Igor Pelgreffi*, «Lo Sguardo», n. 34, 3, 2022.

limits already mentioned – of a proto-sociality or a quasi-sociality, which is not necessarily a pseudo-sociality.

The hypothesis sketched in this article outlines the possible shapes of an *ethics of relationship*, where the relational aspect *precedes* the position of *related values*: an ethics closely linked, in its form, to the co-evolution of agents (natural or artificial) but at the same time an ethic that calls and calls to responsibility towards our *ability to read* the whole of all the emerging dynamics, of the current transformations, of the *uncunny* scenarios of contamination between biological and artificial, at various levels.

This seems to be the great challenge, like a void space of understanding that is now open in front of us, but which perhaps needs to be investigated, by looking for new languages, new vocabularies, new theoretical figures: it will be a slow and patient construction towards something still undetermined. At that point the algorithms could really be thought of as having a life of their own: a *life different* from the one we are used to thinking, for millennia, but still a *form of life*, a *Lebensform* of those who are 'differently alive', of systems equipped with their surprising organization, not entirely programmable, but not for this unmanageable by the human community, from forms not all pre-contained in the initial program. And where the property of 'not being pre-contained' has to do with improvisational skills that also become, in some way still not codified, emotional and social capacities.

In short, the elaboration of an ethics capable of conceiving not only our relationship with the machine but also – in perspective – of the algorithms towards us could have some use in the near future (mixing somehow a top-down approach and a bottom-up approach[65]). But then it will also be necessary to admit that, perhaps, in this context the term ethics is no longer adequate: something different will be needed, which starts from what we currently mean by ethics but is able to proceed further, completing and expanding our ethical vision. Ethics changes its very meaning precisely because it comes into contact with a disturbing or simply enigmatic reality, which it cannot dominate. Perhaps a *synthetic ethics*[66] (Dumouchel and Damiano) or an *incremental ethics*[67]

[65] See also B. Christian, *The Alignment Problem. Machine Leaning and Human Value*, W. W. Norton & Company, 2020; B. Goertzel, *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, «Journal of Artificial General Intelligence», 2014, 5, 1, pp. 1-46.
[66] P. Dumouchel, *Vivre avec les robots*, cit., p. 30.
[67] Similar hypotheses have also been advanced by other scholars, for example I. van de Poel, *An Ethical Framework for Evaluating Experimental Technology*, «Science and Engineering Ethics», 22, 2016, pp. 667-686: F. Amigoni, V. Schiaffonati, *Ethics for Robot as Experimental Technologies. Pairing Anticipation with Exploration to Evaluate the Social Impact of Robotics*, «IEEE Robotics and Automation Magazine», 1, 2018, pp. 30-36. In particular, Viola Schiaffonati recently proposed the idea of an incremental ethics, based on the existence of a close connection between the over-determined nature of some experimental results of digital technologies, in particular of the more 'extreme' and not *explainable* algorithms, and a correlative ethical return characterized by a 'more advanced' approach than traditional ones. Ethics itself becomes less 'static' and more exploratory, in turn, progressive and precisely 'incremental' with respect to the field of observation of the phenomena to be 'ethicized': «the epistemic uncertainty that

(Schiaffonati), in which the ethical aspects are continuously integrated with epistemological or technical philosophy aspects.

All this could involve a complexification of ethics itself, in which it opens its models to greater interaction with epistemology: an *epistemologically oriented* ethics? Ultimately, deconstructing the internal/external, natural/artificial dichotomy is certainly also a risk; but not pursuing it, at least in philosophical reflection, does not involve an even greater risk? And that is the risk of finding oneself, in *automatic society* (Stiegler) or in the *infosphere* (Floridi), totally immersed, involved, subjectivated but also without conceptual tools, inadequate tools as they are 'detached' from the materiality of an experience that could be, in a few years or decades, strongly characterized and crossed by the digital, by the algorithm, by the artificial autonomy, albeit in the primitive stages of a slow evolution?

Igor Pelgreffi
Università degli studi di Verona
✉ igor.pelgreffi@univr.it

References

Primary Bibliography

Anderson M., S. Anderson (eds.) 2014. *Machine Ethics*, Cambridge University Press, Cambridge.

Aristotle 2000, *Nicomachean Ethics*, tr. by R. Crisp, Cambridge University Press, Cambridge.

Bateson G. 1972. *Steps to an ecology of mind*, Ballantine Books, New York.

Berardi F., Sarti A. 2008. *Run. Forma, vita, ricombinazione*, Mimesis, Milano-Udine.

Bostrom N. 2014. *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Canguilhem G. 1992. *Machine and Organism* [1952], in «Incorporations» (J. Crary, S. Kwinter eds.), Zone Books, New York, pp. 45-69.

Christian B. 2020. *The Alignment Problem. Machine Leaning and Human Value*, W. W. Norton & Company.

D. A. Norman 2005. *Emotional Design*, Basic Books, New York.

Damasio A. 2004, *Looking for Spinoza. Joy, Sorrow and the Feeling Brain*, HMH, Montreal.

lies at the heart of exploratory experimentation can be translated from an ethical point of view in a form of incrementalism, where exploratory experiments play a role precisely in acquiring knowledge on how these experimental technologies operate in their contexts of use» (cf. V. Schiaffonati, *Computer, robot ed esperimenti*, «aut aut», 392, 2021, pp. 51-62, p. 59).

Derrida J. 1994. *Nietzsche and the machine*, «Journal of Nietzsche Studies», 7, pp. 7-66.
Dumouchel P., Damiano L. 2016. *Vivre avec les robots. Essai sur l'empatie artificielle*, Seuil, Paris.
Fadini U., 1999, *Principio metamorfosi. Verso un'antropologia dell'artificiale*, Milano, Mimesis.
Guattari F. 1995. *Chaosmosis. An ethico-aesthetic paradigm* [1992], Indiana University Press, Indiana.
High-Level Expert Group on AI 2019. *Ethics Guidelines for Trustworthy AI*, EU Guidelines, Bruxelles.
Houdé O. 2017. *Apprendre à résister*, Le Pommier, Paris.
Lin P., Jenkins R., Abney K. 2017. *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*, Oxford University Press, New York.
Liu J. et al (eds.) 2005. *Autonomy Oriented Computing. From Problem Solving to Complex Systems Modelling*, Kluwer Academic Publisher, New York.
Mauss M. 1973. *Techniques of the body* (1934), tr. eng. B. Brewster, «Economy and Society», 2, pp. 70-88.
Merleau-Ponty M. 1936, *The Structure of Behaviour*, Beacon, Boston.
OECD 2021. *OECD Council Recommendation on Artificial Intelligence* (2021), OECD, Paris.
Paiva A. 2000. (ed.). *Affective Interactions: Toward a new generation of computer interfaces*, Springer, Berlin.
Picard R. W. 1995. *Affective Computing*, M.I.T. Media Laboratory Perceptual Computing Section, Technical Report n. 321, Boston
Prokopenko M. (ed.) 2014. *Guided Self-Organization. Inception*, Springer, Heidelberg.
Sarti A., Citti G., Piotrowski D. 2022, *Differential Heterogenesis. Mutant forms, sensitive bodies*, Springer, Berlin.
Simondon G. 2014. *Sur la technique*, PUF, Paris.
Sloterdijk P. 2009. *Du mußt dein Leben* ändern. Über Anthropotechnik, Frankfurt 2009.
Tamburrini G. 2000. *Etica delle macchine*, Carocci, Roma.
Thompson E., Varela F. J. 2001. *Radical embodiement. Neural dynamics and conscoiusness*, «Trends in Cognitive Sciences», 5, 10, pp. 418-425.
Tzsafestas S. G. 2016. *Roboethics. A Navigating Overview,* Springer, Berlin.

Secondary Bibliography

Amigoni F., Schiaffonati V. 2018, *Ethics for Robot as Experimental Technologies. Pairing Anticipation with Exploration to Evaluate the Social Impact of Robotics*, «IEEE Robotics and Automation Magazine», 1, pp. 30-36.
Bernacer J., Lombo J. A., Murillo J. I.(eds.) 2015. *Habits: plasticity, learning and freedom*, «Frontiers in Human Neuroscience», Lausanne.

Bostrom N. 2002. *Existential Risks. Analysing Human Extinction Scenarios*, «Journal of Evolution and Tecnhology«, 9.

Bostrom N., Yudkowsky E. 2014. *The Ethics of Artificial Intelligence*, in *The Cambridge Handbook of Artificial Intelligence*, K. Frankish, W. M. Ramsey (eds.), Cambridge University Press, Cambridge, pp. 316-334.

Breazeal C. 2003. *Emotions and sociable humanoid robots*, «International Journal of Human-Computer Studies», 59, 1, pp. 119-155

Breazeal C. 2003. *Toward sociable robots*, «Robotics and Autonomous Systems», 42, 3-4, pp. 167-175.

Camras L. A. 1992. *Expressive development and basic emotions*, «Cognition and Emotion», vol. 6, n. 3 and 4.

Damiano L., Stano P. 2018. *Synthetic biology and artificial intelligence. Grounding a cross-disciplinary approach to the synthetic exploration of (embodied) cognition*, «Complex Systems», 27, 3, pp. 199-228.

Damiano L., Dumouchel P. 2011. *Artificial empaty, imitation and mimesis*, «Ars vivendi», 1, 1, pp. 18-31.

Damiano L., Dumouchel P., Lehmann H. 2015. *Towards human-robot affective coevolution. Overcoming oppositions in constructing emotions and empathy*, «International Journal of Social Robotics», 7, 1, pp. 7-18.

Dautenhahn K. 2007, *Methodology and themes of human-robot interaction. A growing research field*, «International Journal of Advanced Robotics», 4, 1, pp. 103-108.

Di Paolo E. 2003. *Organismically-insipred robotics. Homeostatic adaptation and teleology beyond the closed sensorimotor loop*, in Murase K., Akasura T. (eds.), *Dynamic Systems Approach for Embodiment and Sociality*, «Advanced Knowledge International», Adelaide, pp. 19-42

Doshi-Velez F., Kim B. 2017. *Towards a rigorous Science of Interpretable Machine Learning*, on line arXiv:1702.08608v2.

Dumouchel P. 1997. *Emotions. Essai sur le corps et sur le social*, «R. Ph. de la France et de l'Etranger», 187, (1).

Esposito E. 2021. *Dall'intelligenza artificiale alla comunicazione artificiale*, «aut aut», 392, pp. 20-34.

EU 2020. *White Paper on Artificial Intelligence. A European approach to excellence and trust*, Bruxelles.

Fabbris, L. 2021. *Il programma, la rete, il circolo: un approccio Io-fi all'Autonomia artificiale*, «aut aut», 392, pp. 63-76.

Floridi L. 2022, *On algorithms: ethical and epistemological questions. Interview with Igor Pelgreffi*, «Lo Sguardo», n. 34, 3.

Fong T., Nourbakhsh I., Dautenhann K. 2003. *A survey of socially interactive robots*, «Robotics and Autonomous Systems», 42, 3-4, pp. 146-166.

Fossa F., Schiaffonati V., Tamburrini G. 2021. *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Carocci, Roma.

Froese T. 2007. *On the role of AI in the ongoing paradigm shift within the cognitive sciences*, in M. Lungarella et al (eds.), *50 Years of Artificial Intelligence, Lecture Notes in Computer Science*, vol. 4850, Springer, Berlin, pp. 63-75.
Goertzel B. 2014. *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, «Journal of Artificial General Intelligence», 5, 1, pp. 1-46.
Houdé O. 2017. *La neuroéducation: magie ou science? Cerveau & Psycho/Pour la science*, «Chronique. L'école des cerveaux», 86, pp. 80-83.
Ienca A., Vayena E. 2019. *The global Landscape of AI Ethics Guidelines*, «Nature Machine Intelligence», 1, 9, pp. 389-399.
Ishiguro K 2021. *Klara and the sun*, Random House Large Print, London.
Lawless W. F. et al (eds.) 2017. *Autonomy and Artificial Intelligence. A Threat or a Savior?*, Springer, Cham.
Li H. *et al*. 2011. *Towards an effective design of social robots*, «Int. J. of Social Robotics», 3, 4, pp. 333-335.
Miller T. 2019. *Explanation in Artificial Intelligence. Insights from the Social Sciences*, «Artificial Intelligence», 267, pp. 1-38.
Mori M. 1970. *Bukimi no tami (The uncanney valley)*, «Energy», 7, 4, pp. 33-35.
Murray I. R., Arnott J. L. 1993. *Toward the simulation of emotion in synthetic speech*, «J. Acoust. Soc. Am.», vol. 93, feb., pp. 1097-1108.
Norman D. A. 2004. *Emotional Machines* in Id., *Emotional design*, Basics Books, New York.
OECD 2020. *Principles on AI OECD*, Paris.
Parisi D. 2014. *Future robots. Towards a robotics science of human beings*, John Benjamins, London.
Parisi D. 2014. *Internal robotics*, «Connections Science», 16,4, pp. 325-338
Pelgreffi I. 2019. *Bernard Stiegler e la critica della società automatica*, in B. Stiegler, *La società* automatica, Meltemi, Milano, pp. 11-26.
Pelgreffi I. 2020. *Ambiente digitale, automatismi e corporeità*, in (F. Miano, L. Alici eds.) *L'etica nel futuro*. Orthotes, Napoli-Salerno, pp. 389-399.
Pfeifer R., Bongard J. C. 2006, *How the body shapes the way of we think*, MITPress, Cambdridge.
Pfeifer R., Scheier C. 1999. *Understanding Intelligence*, MIT Press, Cambridge MA.
Pitt L., Valiant L. 1988. *Computational limitations on learning from examples*, «Journal of the ACM», 35, pp. 965-984.
Ramus F. 2018. *Neuroéducation et neuropsychanalyse: du neuroenchantement aux neurofoutaises*, «Intellectica», 1-2, 69, pp. 289-301.
Ryan M., Stahl B. 2021. *Artificial Intelligence Guidelines for Developers and Users*, «Journal of Information, Communication and Ethics in Society», 19, 1, pp. 61-86.
Schiaffonati V. 2021. *Computer, robot ed esperimenti*, «aut aut», 392, pp. 51-62.
Schmetkamp S. 2020, *Understanding AI. Can and Should we Empathize with Robots?*, «Review of Philosophy and Psychology», 11, pp. 881-887.
Scott R. 1982, *Blade Runner*, USA.

Sekuper R., Blake R. 1998. *Star Trek on the Brain. Alien Minds, Human Minds*, Freeman, New York.

Sharkey A., Sharkey N. 2010. *The crying shame of robot nannies. An ethical appraisal*, «Interaction Studies», 11, 2, pp. 161-190

Sharkey A., Sharkey N. 2012. *Granny and the robots. Ethical issues in robot care for elderly*, «Ethics and Information Technology», 14, 2, pp. 27-40.

Sparrow R., Sparrow L. 2006. *In the hand of machines? The future of aged care*, «Minds and Machines», 16 , 2, pp. 141-161.

Testa I., Caruana F. (eds.) 2020. *Habits. Pragmatist Approaches from Cognitive Science, Neuroscience, and Social Theory*, Cambridge University Press, Cambridge.

Tisseron S. 2015. *Le jour où mon robot m'aimera. Vers l'empathie artificielle*, Albin Michel, Paris.

Vallor S., Bekey G. A. 2017. *Artificial Intelligence and the Ethics of Self-Learning Robots*, in Lin et al (eds), *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*, Oxford University Press, New York, pp. 338-353.

van de Poel I. 2016. *An Ethical Framework for Evaluating Experimental Technology*, «Science and Engineering Ethics», 22, pp. 667-686.

Veruggio G. 2007. *Euron Roboethic Roadmap*, Release 1.2, january.

Veruggio G., Operto F., Bekey G. 2017. *Roboethics. Social and Ethical Implications*, in Siciliano, Khatib, *Springer Handbook of Robotics*, Springer, Berlin[2], pp. 2135-60.

Wallach W., Allen C. 2009. *Moral Machines. Teaching Robots Right from Wrong*, Oxford University Press, Oxford

Weinberger D. 2017. *Machines Now Have Knowledge We'll Never Understand*, «Wired», on line 18.4.2017.

Whatcher S., Mittelstadt B., Floridi L. 2017. *Transaparent, Explainable and Accountable AI for Robotics*, «Science Robotics», 6.

Ziemke T. 2003, *What's that thing called embodiement?*, in *Processing of the 25th Annual Meeting of the Cognitive Science Society*, Lawerence Erlbaum, pp. 1305-1310.

Ziemke T. 2016. *The body of knowledge*, in L. Damiano, Y. Kuruma, P. Stano (eds.), *What can shyntetic biology offer to artificial intelligence (and viceversa)*?, «Byosystems», 148, pp. 4-11.

Ziemke T., Lowe R. 2009. *On the role of emotion in embodied cognitive architectures: from organisms to robots*, in «Cognitive Computation», 1, 1, pp. 104-107.