Deep Learning Opacity, and the Ethical Accountability of AI Systems. A New Perspective

Gianfranco Basti and Giuseppe Vitiello

Abstract In this paper we analyse the conditions for attributing to AI autonomous systems the ontological status of "artificial moral agents", in the context of the "distributed responsibility" between humans and machines in Machine Ethics (ME). In order to address the fundamental issue in ME of the unavoidable "opacity" of their decisions with ethical/legal relevance, we start from the neuroethical evidence in cognitive science. In humans, the "transparency" and then the "ethical accountability" of their actions as responsible moral agents is not in contradiction with the unavoidable "opacity" (unawareness) of the brain process by which they perform their moral judgements on the right action to execute. In fact, the moral accountability of our actions depends on what is immediately *before* and *after* our "moral judgements" on the right action to execute (formally, deontic first order logic (FOL) decisions). I.e., our moral accountability depends on the "ethical constraints" we imposed to our judgement *before* performing it in an opaque way. Anyway, our moral accountability depends overall on the "ethical assessment" or explicit "moral reasoning" *after and over* the moral judgement before executing our actions (deontic higher order logic (HOL) assessment). In this way, in the light of the AI "imitation game", the consistent attribution of the status of ethically accountable artificial moral agents to autonomous AI systems depends on two similar conditions. Firstly, it depends on the presence of "ethical constraints" to be satisfied in their Machine Learning (ML) supervised optimization algorithm during its training phase, to give the system ethical skills ("competences") in its decisions. Secondly – and definitely—, it depends on the presence in an AI autonomous system of a deontic HOL "ethical reasoner" to perform an automatic, and fully transparent assessment (metalogical deontic valuation) about the decisions taken by the ethically skilled ML algorithm about the right action to execute, before executing it. Finally, we show that the proper deontic FOL and HOL for this class of artificial moral agents is Kripke's modal relational

G. Basti
Pontifical Lateran University, Vatican City, Italy
e-mail: basti@pul.va

G. Vitiello (✉)
Department of Physics "E.R. Caianello", University of Salerno, Fisciano, Italy
e-mail: gvitiello@unisa.it

R. Giovagnoli and R. Lowe (eds.), *The Logic of Social Practices II*, Studies in Applied Philosophy, Epistemology and Rational Ethics 68,
https://doi.org/10.1007/978-3-031-39113-2_2

logic, in its algebraic topological formalization. This is naturally implemented in the dissipative QFT unsupervised deep learning of our brains, based on the "doubling of the degrees of freedom" (DDF), and then in the so-called "deep-belief" artificial neural networks for the statistical data pre-processing. This unsupervised learning procedure is also compliant with the usage of the "maximin fairness principle", used as a balancing aggregation principle of the statistical variables in Sen's formal theory of fairness.

Keywords Machine ethics · Deep learning opacity · Higher order deontic reasoners · Relation ethics · Biased statistical data · Deep-belief neural networks · Fair machine learning

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
BAO	Boolean Algebra with Operators
DDF	Doubling of the Degrees of Freedom
FOL	First-Order Logic
HOL	Higher-Order Logic
ME	Machine Ethics
ML	Machine Learning
NE	Neuroethics
NWF	Non-wellfounded set theory
QFT	Quantum Field Theory
QM	Quantum Mechanics
SCT	Social Choice Theory
TCS	Theoretical Computer Science
TM	Turing Machine
TPM	Transitive Probability Matrix
UTM	Universal Turing Machine

1 Introduction: The Ethics of Artificial Intelligence and the Machine Ethics

1.1 The Distributed Responsibility Humans-Machines in Artificial Intelligence

Recently, L. Floridi and M. Taddeo introduced into the wide debate about ethics in AI the notion of *distributed responsibility* between humans (designers, developers, users), on the one hand, and machines (software and hardware), on the other hand:

> The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware. This is known as distributed agency. With distributed agency comes *distributed responsibility*. [1, p. 751]

More recently, one of us proposed to redefine the notion of "distributed responsibility" between humans and machines, by distinguishing between the *slow responsibility* of conscious ethical agents such as humans, and the *fast responsiveness* of unconscious skilled moral agents such as machines with respect to the ethical constraints from the shared social environment [2]. This distinction extends to *Machine Ethics* (ME) a similar distinction used already in *Neuroethics* (NE) as to the relationship between the slowness of consciousness with respect to the fastness of the related neural processes of the *human person agency* [3, 4]. Indeed, it is the *person* (i.e., the individual-in-relationship with her physical-social environment) and neither her mind, nor her brain taken in isolation *the proper subject* of morally/legally accountable actions [5, 6] (see Sect. 2.1).

Effectively, the main problem at issue about the actual challenges in moral philosophy related to NE, and *Artificial Intelligence* (AI) is similar. "Who is the actor of a moral act?" or, using the title of a famous M. S. Gazzaniga's book: "Who is in charge" [7]. Now, as N. Levy[1] emphasized, what NE teaches us is that neither the conscious mind of a human person, nor some part of her brain such as the lobes of the prefrontal cortex, as far as both taken in isolation, can be considered as the "controller" of the human behavior.

> We needn't fear that giving up on a central controller requires us to give up on agency, rationality, or morality. We rightly want our actions and thoughts to be controlled by an agent, by ourselves, and we want ourselves to have the qualities we prize. But the only thing in the mind/brain that answers to the description of an agent is the entire ensemble: built up out of various modules and sub-personal mechanisms. And it is indeed the entire agent that is the controller of controlled processes. [8, p. 41]

In this sense, the same N. Levy suggests that AI systems, as far as considered as artificial extensions of the human natural intelligence can be considered among these sub-personal mechanisms and modules, in the framework of the so-called *extended mind* hypothesis in cognitive neurosciences (see [8, pp. 29–44]).

1.2 AI Subjects of Moral Agency and the Machine Ethics

However, V. C. Müller well emphasizes in his recent review paper on *Ethics of Artificial Intelligence and Robotics* [9] that considering AI systems as simple extensions of the human intelligence covers only one trend of the actual debate about the

[1] Neil Levy is professor at the University of Oxford where he is also Director of *The Oxford Uehiro Centre for Practical Ethics* that has in "Neuroethics" and in "AI and Digital Ethics" two of the main topics of research. We emphasize in this paper the strict relationship between these two research fields, because both have as object the physical bases of ethics (effectively, of deontic logic information processing) respectively in natural and artificial neural systems.

ethics in AI. Namely, the trend concerning the AI systems as *objects,* that is, as tools made and used by humans (individuals, companies, private and public institutions, etc.) as indispensable support of the human decision-making for the management of extremely large bases of data ("big-data").

The other trend that is more relevant for the ME debate concerns the AI systems as *subjects* and then as *artificial moral agents.* I.e., as *autonomous systems* able to make *fast* decisions escaping the *slow conscious* human control but affecting our individual and social lives and then with evident ethical/legal consequences of their decisions. Evidently, these artificial moral agents require that they satisfy a suitable *ethical/legal accountability* of their decisions just in the way it happens for the human moral agents.

Therefore, following Müller's useful distinction, we have

> Ethical issues that arise with AI systems as *objects*, i.e., tools made and used by humans. This includes issues of privacy and manipulation, opacity and bias, human-robot interaction, automation and employment, and the effects of autonomy. Then AI systems [can be considered also] as *subjects* requiring an ethics for the AI systems themselves in machine ethics and artificial moral agency. [9, p. 1]

More analytically the main ethical issues in AI concern:

1. *Issues about privacy and data manipulation.* They are the more discussed and easier to be understood. AI systems are applied wherever there are large databases ("big-data") whose management is impossible for humans, and which now with the progressive informatization of any aspect of our personal, social, and economic lives, concern the sensitive data of all of us.
    - What perhaps escapes most and it is paradoxical but true, is that these systems, by profiling us and cross-relating the data concerning us every time we use Internet or our smartphones, make an online purchase, access a database, request an online document, or simply we use an internet search engine, they know our habits, attitudes and preferences *much better than we know ourselves.*
    - These profiles are *accessible to others and not to us,* which creates a big ethical-legal problem that we should sooner or later face as individuals and as governments. In fact, these profiles are used systematically in *the creation of fakes* to influence specific groups of people, with serious problems on the autonomy of choices not only in the economic-commercial field, but also in the *political-social field*.
    - Representative democracies like ours no longer work if the citizen choices are *systematically conditioned* in a subtle but real way. Burying our heads in the sand as we are doing does not solve the problem but exacerbates it. And this constitutes a real problem for Western democracies in that *undeclared but effective war* between democratic and autocratic regimes, in which we are all sadly involved since many years (see [10] for further discussions).
2. *Issues about opacity and bias in the statistical data processing.* As it is well known, the classic *expert systems* in the automatic processing and classification of

data related to the so-called "*symbolic* approach to AI" (see Sect. B2 in Appendix B) do not suffer from this kind of problems. On the contrary, the much more powerful AI systems that include *machine learning* (ML) algorithms based on multilayer architectures of neural networks (the so-called "deep-learning": see Sects. B4–B7 in Appendix B) systematically suffer from an unavoidable problem of *opacity* in data processing. Because of their relevance, these two strictly related issues are the main object of our paper (see Sects. 2 and 3).

- In expert systems of the symbolic AI, indeed, the inferential trees for data classification are defined by the programmer and therefore the path followed by the system to reach the final decision *can be always reconstructed* and therefore it is controllable, or "transparent".
- This is *systematically impossible* in ML models based on multilayer neural networks, which moreover *necessarily emphasize "biases"* or "negative propensions" towards certain groups or types of individuals—generally minorities—eventually present in the statistical data on which the training of the system is carried out. This raises "significant concerns about lack of due process, accountability, community engagement and auditing" in AI systems for automated decision support [11, p. 18ff] and it requires the necessity of inserting into the *unsupervised pre-processing* of the training set in ML algorithms *fairness criteria* to correct these distortions, as we discuss below (see Sects. 3.2, B7 in Appendix B, and Sect. D2 in Appendix D).

3. *Issues about the human–robot interaction.* Although still not too obvious to many compared to the previous problems, it is an emerging ethical-legal issue, which will become increasingly relevant, as robot usage will be spread on a very large scale.

   - Robots—including *self-driving* aerial and ground vehicles—are indeed destined to support or even replace humans in *industry*, *communications, services* (e.g., automatic call centers), *surgery* (surgical robots), *high-risk rescue operations,* and increasingly in *military operations* (armed drones, robot-soldiers, robotic artillery, etc.), all specific fields where they are already widespread. Their usage will increase also in many other applications that affect the lives of all of us (think at self-driving cars), even the more *fragile*. On this regard, think at robot applications in the *nursing care*, in the *domestic care*, and even in the *educational care* (i.e., the distance teaching systems endowed with AI engines for readapting themselves to the individual student needs).

4. *Issues related to autonomous systems and the ME.* It is evident that the discussion about AI systems as "subjects of moral agency" and then as "artificial moral agents" in ME concerns essentially AI systems, as far as displaying different degrees of *autonomy* in their decision making, with respect to the human control. As Müller properly recalls,

> There are several notions of autonomy in the discussion of autonomous systems. *A stronger notion* is involved in philosophical debates where autonomy is the basis for responsibility and personhood [12]. In this context, responsibility implies autonomy, but not inversely, so there can be systems that have degrees of technical autonomy without raising issues of responsibility. The weaker, more technical, notion of autonomy in robotics is relative and gradual. A system is said to be autonomous with respect to human control to a certain degree. There is a parallel here to the issues of bias and opacity in AI since autonomy also concerns a power-relation: who is in control, and who is responsible? [9, pp. 24–25]

- The examples made by Müller of AI autonomous systems about which the moral issues are object of fierce debates, both from a technical point of view, and from an ethical and juridical perspective are the "self-driving cars" and the "autonomous weapon systems" (AWS, e.g., armed drones and robot-soldiers) (see [9, pp. 24–29] and the quoted literature about these topics). They are examples of high relevance and actuality for our society.
- Of course, ME—namely, "the ethics for machines as *subjects*, rather than for the human use of machines as *objects*"—is strictly related to the issues of autonomy and opacity in AI systems. That is, using a quotation of V. Dignum in Müller's paper [9, p. 30], ME is concerned with the ambitious constraints that:

> AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. [13, pp. 1–2]

1.3 A Scheme of This Contribution

The last two quotations of Mülller and Dignum help us to define better what we mean by the autonomy of AI systems intended as *subjects of moral agency.* Then they help us for defining properly the notion of *distributed moral responsibility* between humans and machines in terms of the *distributed ethical/legal accountability for the rest of the society* of their decisions/actions because concerning the life and welfare of persons. To sum up, in the case of autonomous AI systems, we must speak about the distributed ethical/legal accountability between *conscious moral subjects* and *artificial moral subjects.*[2]

[2] In this connection, it is useful to recall that the notion of *responsibility/accountability* in moral philosophy is also etymologically related to the notion of "responding/accounting to someone else" for our actions. Or, in the formal terms of the *deontic logic*, the notion of responsibility consists in *justifying the consistency* of our actions with respect to the *obligations* of norms ruling our behaviors, for satisfying a given system of *values.* I.e., a *heterarchy* (= a hierarchy in which the ordering can change [115]) *of individual/common goods* to be pursued which are shared by the members of a given community. In this sense, in the computational implementation of deontic logic calculations in autonomous AI systems making them *artificial moral agents*, it makes sense speaking about the

As a premise, it is significant for our aims Müller's distinction between the "philosophical"—effectively, *anthropological*—and the "technical" notions of *autonomy* and *control* in humans and machines.

On the one hand, autonomy and control are essential components of the human personhood. Indeed, we can define the *personal free-will* of humans as "the capability of a human person of controlling at different levels her own behavior, in view of the effective pursuing of a given goal (value) by suitable decisions/actions" (see [5], Chap. 5) and/or in Amartya Sen's terms "in view of the effective pursuing of a valued and valuable state/way of living" [14, p. 356].

On the other hand, the "relative and gradual" autonomy of AI systems/robots with respect to the human control is strictly related to what Müller defines as the "technical" notion of control in artificial systems. This effectively refers to the basic notions of *Cybernetics,* or "(Theory of) Communication and Control in Animals and Machines", according to the title of famous Norbert Wiener's book [15]. In Sect. A1 of the Appendix A we recall briefly which are the *three main levels* of active control in biological and artificial systems, emphasizing that the autonomy of AI Systems with respect to the human control reaches its higher level when it concerns the same ultimate level of the goals supervising the behavior. This is strictly related to the implementation into the decision processes of AI autonomous systems—before all in the ML optimization process—of deontic logic constraints as necessary condition for attributing them the ontological status of artificial moral agents (see Sect. A2 in Appendix A).

Therefore, main object of this contribution is a theoretical justification of the attribution of the *ontological status* of artificial moral agents to autonomous AI systems in ME by a systematic comparison, in the light of the Turing Test, to the human persons as conscious moral agents.

More precisely, an adequate *connotation* or "descriptive definition" of the autonomous AI systems that justifies, in a logically and ontologically consistent way, our *reference* to them, as artificial moral agents, requires the fulfilment of the following steps that we will examine in the rest of the article. Even though, for not burdening our discussion, the steps 1–3 that refer to the background knowledge necessary for defining the notion of "autonomous AI systems" will be treated in the Appendix A and in the Appendix B of this paper.

1. A connotation of the ethical decision-making autonomy of humans and machines within the already introduced notion of *active control,* which is what distinguishes biological and artificial systems from mechanical systems as *passive control* systems (see Appendix A).
2. A connotation of the ethical decision-making autonomy of humans and machines in the framework of the *cognitive science triangulation* among (a) the *intentional states* of the subjective mind, as such inaccessible to other people; (b) the *physical (neurophysiological) states* to which they are necessarily related; (c) the *logical operations* implementable in (b), which expresses (behaviorally/linguistically)

satisfaction of ethical constraints from the social environment they share with humans (see Sect. 2 and Appendix A for more details).

the "intelligence" both cognitive and moral of (a), and hence makes it imitable by suitable AI system models (see Sect. B1 in Appendix B).

3. A connotation of the ethical decision-making autonomy of humans and machines in the context of the *Turing imitation test* that is at the origin of the AI research program. With the consequent distinction between: (a) the *symbolic AI systems* or "expert systems" because they simulate in the explicit inference tree of a program the ability of a human expert (see Sect. B2 in Appendix B); and (b) The *pre-symbolic AI systems* because they are equipped with ML algorithms based on different models of multilayered ANNs (*deep-learning*). Indeed, the vastness of the databases (*big-data*) on which they apply excludes its treatability by any human expert, making these systems *indispensable* to our society (see Sects. B3–B7 in Appendix B).

Now, precisely this class of autonomous AI systems endowed with multilayered ML models display the problems of *opacity* and *unfairness* in their decision processes that, as we anticipated in Sect. 1.2, are the two main issues to solve in ME (see item 2 in the list). Therefore, in the next two Sections of this work, these two problems and their possible solutions in ME are discussed. In other words, the proper attribution of a *subjective moral agency* to autonomous AI systems in comparison with the moral agency of the human subjects requires the fulfillment of these further two steps:

4. A connotation of the ethical decision-making autonomy of humans and machines, and of their accountability, despite the intrinsic *opacity* of the decision processes in both. In other terms, the two conditions of *transparency* that the human moral agents must satisfy to grant the ethical/legal accountability of their decisions/actions when they concern the life and the welfare of other persons, despite the intrinsic opacity of the human decision processes must be satisfied also by autonomous AI systems in similar conditions. To the illustration of this fundamental opacity issue and of its solution both in humans and machines is devoted the Second Section of this paper.
5. A connotation of the ethical decision-making autonomy of humans and machines as able to perform *fair* moral judgements/decisions, so overcoming the discriminations towards persons and groups present in the linguistic/social/cultural environments of which they are parts. To this issue in ML models and to its possible solution according to Amartya Sen's theory of "justice as fairness" we dedicate the Sect. 3 of our paper and the related Appendixes C and D.

2 The Autonomy and the Opacity Issues in Machine Ethics and the "Imitation Game" in the AI Research Program

2.1 The Contribution of Neuroethics to Solve the Problem

In this section we discuss mainly the issue of how granting *the ethical/legal right to transparency* with respect to the decisions of autonomous AI systems if this transparency is technically impossible in these systems.

To solve this typical conundrum of the "distributed responsibility" humans-machines, the extension to AI systems of the neuroethical distinction between the *fast unconscious responsiveness* of our brains to environmental constraints and the *slow conscious responsibility* of our minds becomes essential.[3] For this aim, it is useful to report here the following example used by N. Levy in his book about the relationship between consciousness and moral responsibility in cognitive neurosciences, in the light of neuroethics (NE) [6]. The example concerns the issue of the moral responsibility of a *highly skilled*—very well "trained", indeed—human driver such as the famous racing driver Ayrton Senna.

> It is characteristic of conscious processes *that they are much slower than nonconscious; the rapid responsiveness of highly skilled agents* like (…) Senna must certainly be driven by the latter and not the former. *It therefore seems false that agents must be conscious of the information they respond to in order to be responsible for how they respond to it.* (…) Direct moral responsibility requires that a creature conscious agent *be conscious of the moral significance of their actions.* [6, pp. 114–121] (Italics are ours)

In this way, Levy introduces into the neuroethical debate the fundamental distinction between the moral responsibility of conscious communication agents such as human persons, versus the fast responsiveness of their brains. But also, versus the (much faster) responsiveness of AI decision supports we use, which can be both (brains and machines) included in the category of sub-personal modules of actions for which we, as persons, are morally responsible.

This distinction, however, can also shed light on the issue of *autonomous AI systems* equipped with ML, understood not as objects but as *unconscious subjects of morally relevant decisions*, or more precisely as *highly trained artificial moral agents*, through ML techniques—think at self-driving cars—, destined to become a significant part of our society.

In other words, in his book Levy emphasizes that in the human production both of *cognitive and of moral judgements*—formally, logical valuations/decisions "true/false" (1/0), in *alethic and deontic* modal logics, respectively (see Sect. A2 in Appendix A)—what is "transparent", i.e., *conscious* to us and eventually "transparent" and then "accountable" to others, is what is *before and after* the production of the judgement (decision) itself that as such is absolutely *unconscious* and then

[3] The implicit reference is to the neurophysiological evidence that the action potentials of the neuron circuits involved in an intentional decision/action in human brains reach their maximum some till some tenths of seconds before the conscious component of an intentional state (see [3]).

"opaque" to everybody, just as it happens to autonomous AI systems endowed with deep ML.

In this sense, we can say that these systems with their intrinsic opacity are "the winners" of the "imitation game" [16] on which the AI research program is based since its origins in 1956 [17], much more than the symbolic AI systems, whose decision process is generally "transparent".

Indeed, when we produce a moral judgement about an action and/or a choice that we must execute, what is "conscious" ("transparent" to us, and eventually to others as far as we communicate it) is only what *precedes and follows* the moral judgement/decision as such. Indeed, we can distinguish three steps in any human moral judgement/decision (see [5], Chap. 5 for more details):

1. At first, we *consciously* examine the different components of the action/choice that we are going to evaluate by a moral judgement over it. That is, we consider mainly the past similar situations, the actual concrete situation, the future practical consequences of our action/choice, and of course also the abstract moral norms that should rule our action.
2. Afterward, by combining through an *unconscious process* these and other components not considered at the first step (before all *emotions,* as NE taught us [18]), we produce our moral judgement (i.e., we make our deontic first order (FO) evaluation) about the action/choice we want to execute.

However, being truly responsible of the moral significance of our actions requires that, before executing our action/choice,

1. As a third step, we make *consciously* a sort of "moral auditing to ourselves" about our moral judgement (i.e., we perform a deontic higher order logic (HOL) reasoning/assessment) for evaluating whether effectively this judgement/decision (deontic first order logic (FOL) decision) we produced about the right action to execute, satisfies all the moral constraints we imposed to it—and eventually other moral constraints we did not consider. Our moral responsibility becomes in such a way an act of "transparent" moral accountability for justifying/explaining also to others and not only to ourselves the morality of our decisions/choices. Formally, this "moral auditing" of our moral judgement/decision about the action to execute is a "valuation of our valuation", that is, it consists into a *metalogical evaluation of our deontic FOL decision process,* requiring formally a *deontic HOL* (see [19] and the Conclusions of this contribution).[4]

[4] Formally, only Kripke's *relational semantics* in mathematical modal logic in its algebraic (topological) interpretation of a *modal Boolean algebra with operators* (modal BAO) (see Sect. D1 in Appendix D) admits both FOL *local* semantics, and HOL *total* semantics, given that Boolean logic is the only "guarded decidable fragment" of FOL. Indeed, the semantics of a Boolean algebra requires *partially ordered* sets defined on a topological *Stone space* [55], and then it can be defined also on *Non-wellfounded (NWF) sets* [32, 62]. In them, no *total set ordering* is admitted but several set trees of partially ordered sets *sharing the same root* (see [89] for further details). The usage of NWF-sets is of course compliant for implementing models of deontic logics for a pluralistic society like ours, characterized by different partial orderings of ethical values (see Sect. A2 in Appendix A).

In this way, we, as conscious moral agents, can fully satisfy the moral/legal accountability to the social community of our decisions/choices/actions, because only at this third step we are able of formally justifying/explaining the morality/legality of our decisions/actions/choices. At the same time, we also satisfy in this way *the right-to-transparency* to other human subjects, whose lives are eventually influenced by our decisions/choices/actions.

2.2 *The Double Condition to Satisfy for a Consistent Attribution of an Accountable Moral Agency to AI Systems in Machine Ethics*

To sum up, the solution of the fundamental issue in ME of how *attributing consistently moral agency* to artificial unconscious agents, and specifically to highly trained AI systems requires that they satisfy *two conditions*:

1. The "transparent" implementation in the supervised ML algorithms of *ethical/legal constraints*. That is, error minimization algorithms satisfying also ethical conditions [20]. In this sense, the so-called "consequentialist" or "value based" approach to deontic logic [21] seems to be more suitable for being directly implemented in ML algorithms since in both cases a cost function is to be minimized,[5] than the so called "virtue ethics" approach [22]. Indeed, the value ethics is also denoted as consequentialist because formally satisfying the following modal logic scheme: "if you *want* to pursue factually this goal (value), you *must* do this". For instance, in the case of AI autonomous systems for trading in the financial markets—now covering a larger part of fast trading operations (about 40%) all over the world—a "good" ML algorithm for trading means that it is not based only on the maximization of profit, but also on the satisfaction of given ethical clauses (e.g., investments not deriving from illegal origins, not based on the exploitation of the workers, etc.). Finally, the value-based deontic logic is compliant also with the implementation of "fairness conditions" in the data pre-processing by an *unsupervised* ML, for avoiding the unwanted "bias" in the training data set of supervised ML, leading to "unfair" decisions of the ML algorithm based on biased data (see [23, 24] and below Sect. 3.2).
2. The implementation in autonomous AI systems of *an automatic ethical/legal auditing* to check in a transparent way whether the decisions taken by the system effectively meet the ethical criteria set in the ML algorithm. And/or, in the case of symbolic AI systems, the ethical criteria implemented in the decision tree of the program. Only recently the researchers in AI started to study this fundamental

[5] On this regard, see below Sect. A2 in Appendix A, where it is evident that the ethical obligation in a "value ethics" requires the satisfaction of an optimization (maximization) condition.

component of ME, requiring a HOL for a metalogical valuation of the effectiveness of the deontic logic algorithms implemented in the ML program and/or in the inferential tree of symbolic AI systems (see [19] and below Sect. 2.3).

2.3 *The Implementation of an Automatic Ethical Auditing in AI Systems*

As we have seen in Sect. 2.1 the ethical/legal auditing just recalled is indeed the way in which the human moral/legal agents satisfy the *right to transparency* of the other human subjects and of the whole society as to our decisions, given the unavoidable opacity characterizing the human minds in taking their decisions. Or, in other terms, this self-auditing ("ethical reasoning" as distinguished from the "ethical judgement or decision") for justifying the morality/legality of our decision is the way by which we as human persons satisfy the *obligation to accountability* of our decisions/actions to everybody, ourselves included, making us fully *responsible moral/legal agents.* Precisely this *transparent automatic self-auditing* is what till now was lacking to autonomous AI systems for fully justifying the ethical accountability of their decisions and then their definition as *artificial moral agents* in ME.

From the formal standpoint of the *Theoretical Computer Science* (TCS), the implementation of an *automatic ethical/legal auditing* of the decisions taken by an AI system requires a *deontic HOL* for performing *the metalinguistic analysis of deontic consistency of the decisions*. This requires the support of *automatic demonstrators* specialized in tasks of deontic logic of higher order than the first [19].

Although unknown to the public because they are used almost exclusively in the field of pure and applied logical and mathematical research, systems of this type exist since many years for the metalogical analysis using a HOL of FOL inferences in logic and mathematics.

These HOL systems are widely used for demonstrating/analyzing particularly complex logical/mathematical theorems [25] but overall, they are applied for the formal consistency analysis of decisions performed by complex automatic control systems that require a *very high degree of robustness to error*. Indeed, because they are used in very delicate fields, an error or inconsistency of the program would have catastrophic consequences, not repairable. These systems therefore perform metacontrols for the *analysis of safety and reliability* of particularly delicate automatic control systems before their commercialization / application in the field.

This analysis in TCS is denoted as *functional programming*, or more exactly, as "functional analysis of programs".

- Think, for instance, at the systems for the *automatic control of the landing of aircrafts by the autopilot.* Effectively, during landing, the pilot always gives the controls of the aircraft to the automatic landing control system managed by the control tower of the airport, so that—with the human support of the air traffic controllers—the pilot remains only for emergency interventions.

- Or think at the programs for the *automatic remote-controlled points in railway stations*; or think at the programs for the *control systems of space travels*, of *nuclear plants*, of *large industrial complexes*, of large (electricity/gas) *power distribution networks*, of large *telecommunication networks*, where the net resources must be allocated and reallocated continuously, according to the degree of occupation of the network at various times and situations, and so on. As we see, these automatic control systems govern ever larger parts of our daily lives and the issue of the automatic continuous check of their reliability/safety/security is fundamental.

Only, recently, due to the tumultuous development of ME problems, the new TCS research sector has opened for the design and engineering of *meta-control systems in the field of ethical reasoners, normative theories, and deontic logics*.

As Christoph Benzmüller of the Freie Universität Berlin—one of the world's leading experts in the field of automatic theorem demonstrators using HOLs [25]—and his colleagues summarized at the beginning of their review article on this new field of research, "the main motivation is the development of appropriate tools for the control and management (governance) of intelligent autonomous systems" [19, p. 1]. Effectively, what they present in their paper is not a model but a development tool for programmers named *LogiKEy—Logic and Knowledge Engineering Framework and Methodology*—"for the design and engineering of ethical reasoners, normative theories and deontic logics". Without entering in further technical specifications, the proposed architecture is based on the semantic incorporation of logical calculations and theories that belong to a (specific) ethical-legal domain, within a classical HOL framework—effectively, Alonso Church's *type theory* [26].

As intuitively represented in Fig. 1, the displayed architecture for an intelligent autonomous system with explicit ethical competency distinguishes the *explicit ethical reasoner* with its *ethico-legal domain theories* (= meta-controller sub-system) from the *AI reasoner/planner* (= controlled sub-system) and from other components. They also include the *application data and knowledge* available to both reasoners (sub-systems). The ethical reasoner takes as input the suggested actions from the AI reasoner/planner and hints to relevant application data and knowledge, as well as to a given ethical-legal domain theory. Then, it produces as output *assessments and judgements* concerning which actions are acceptable or not, and it also provides the corresponding *explanations*. That is, the actions suggested by the AI reasoner (controlled sub-system) in Fig. 1 are not executed immediately, but additionally assessed by the ethical reasoner (meta-controller subsystem) for compliance with respect to the given ethico-legal domain theory. This assessment is intended to provide *an additional, explicit layer of explanation and control* on top of the AI reasoner, which already comes with solid own ethical competency [19, pp. 2–3].

In other terms, deontic HOL systems such as the "explicit ethical reasoner" just outlined fully satisfy the second condition we defined for ethically/legally accountable AI systems interpreted as *artificial moral agents*. What is relevant in the architecture outlined is that its metalinguistic consistency analyses can be applied to different deontic logic models, i.e., to different ethico-legal domains.

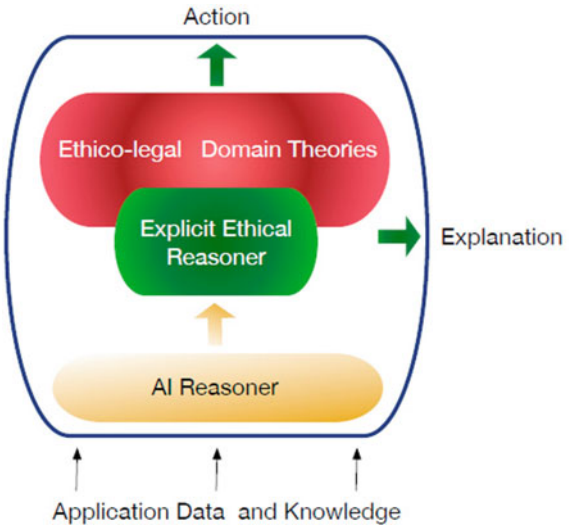

Fig. 1 Schematic representation of the proposed ethical reasoner for the automatic ethical auditing of AI autonomous systems (from [19], p. 3)

Moreover, the autonomous AI system (AI reasoner) whose decisions are the input of the ethical reasoner can be, either of the symbolic type, or of the pre-symbolic one. That is, its own "ethical competency" assessed by the ethical reasoner can be implemented, either as deontic logic algorithms in the decision-tree of its program, or as ethica/legal constraints imposed to the optimization function in which any supervised ML algorithm ultimately reduces itself.

Finally, the ethical reasoner metalogical assessments, not only are fully explicit because no ML algorithm can be implemented in it just as for whichever HOL system (i.e., it is always an AI system of the symbolic type) but it is able to give suitable "explanations" of its assessments over the ethical/legal evaluations of the AI reasoner under scrutiny. In this way, it *fully satisfies* the "right to transparency" that the society must pretend from the autonomous AI systems.

3 Relation Ethics and the Fairness Issue in Machine Ethics

3.1 Relation Ethics and Fairness in a "Liquid Society"

Let us consider now the fifth condition in the list defined in Sect. 1.3 for the consistent attribution of the ontological status of *autonomous moral agent* both to humans and machines. That is, the capability of performing *fair moral judgements/ decisions* despite the discriminations presents toward individual and groups in the social environment of which they are part, and on which their education/training depend.

The necessity of implementing in ethically accountable AI systems fairness ethical criteria to avoid biases in the statistics on which the training phase of a ML model is performed has been recently defined as the necessity of avoiding unintended but real "algorithmic injustices" [27]. Avoiding these problems requires, indeed, "developing and deploying ethical algorithmic systems" satisfying a *relational approach* to ethics.

Effectively, when the standard statistical

> …machine learning systems that infer and predict individual behaviour and action, based on superficial extrapolations, are deployed into the social world, various unintended problems arise. These systems 'pick up' social and historical stereotypes rather than any deep fundamental causal explanations. In the process, individuals and groups, often at the margins of society that fail to fit stereotypical boxes, suffer the undesirable consequences. Various findings illustrate this: bias in detecting skin tones in pedestrians; bias in predictive policing systems; gender bias and discrimination in the display of STEM career ads; racial bias in recidivism algorithms; bias in the politics of search engines; bias and discrimination in medicine; and bias in hiring, to mention but a few. [27, p. 1]

This means that AI algorithms, when applied to automated supports for decision-making processes in the social, political, and economic sphere are not at all "value-free" or "a-moral". The "relational ethics approach" in developing an ethically accountable AI is indeed based on the evidence that "neither people nor the environment, are static; what society deems fair and ethical changes over time" [27, p. 6]. And, we add, "over space too", in the sense that they change for the different groups composing a society. Before all, for the minorities and/or for all the marginalized groups also when numerically consistent, often not having the same welfare opportunities, often sharing different value systems, as well as different criteria of personal flourishing as to the rest of a society. This continuous variability over "space" and "time" of the value systems, of the welfare opportunities and then of the fairness criteria, as well as over the group composition in which the different values systems and welfare opportunities are embedded is indeed what characterizes our *liquid* society and its many legal and ethical issues [28].

3.2 The Relation Ethics in Sen's Theory of Comparative Justice as Fairness and its ML Implementation

Recently, Pratik Gajane and Mykola Pechenizkiy, in a review paper dedicated to the formalization of different fairness criteria in ML algorithms [23], complained the lack of formalization in the ML literature of Amartya Sen's (Economic Sciences Nobel Prize, 1998) approach to *justice as fairness,* in the framework of his *comparative theory of distributive justice*. This lack in ML research occurred despite the relevance of this theory that, for instance, has been used in several documents of the United Nations in the foundations of human development paradigm. What characterizes Sen's approach to fairness, for instance with respect to other approaches identifying fairness with the "equality of opportunities", is that

> variations related to the protected attributes like age, sex, gender, race, caste give individuals unequal powers to achieve goals even when they have the same opportunities. In order to equalize capabilities, people should be compensated for their unequal powers to convert opportunities into "functionings" or "suitable states of being and doing". (…) Crucially, the notion of equality of capability calls for addressing inequalities due to social endowments (e.g. gender) as well as natural endowments (e.g. sex), in contrast to the equality of resources. [23, p. 4]

Effectively, for making some steps in the direction of formalizing Sen's theory of fairness in ML algorithms it is useful to start from its logical formalization in the context of the so-called *social choice theory* (SCT) [14, 29], of which Sen himself was one of the founders,[6] together with another Nobel Prize in Economics, Kenneth Arrow.

> When a group needs to make a decision, we are faced with the problem of *aggregating* the views of the individual members of that group into a single collective view that adequately reflects the "will of the people". How are we supposed to do this? This is a fundamental question of deep philosophical, economic, and political significance that, around the middle of 20th century, has given rise to the field of *Social Choice Theory*. [29, p. 333]

As Sen synthesizes in his Nobel Lecture,

> SCT provides a general approach to the evaluation of, and choice over, alternative social possibilities (including inter alia the assessment of social welfare, inequality, and poverty). (…) If there is a central question that can be seen as the motivating issue that inspires social choice theory, it is this: how can it be possible to arrive at *cogent aggregative judgments* about the society (for example, about "social welfare", or "the public interest", or "aggregate poverty"), given the diversity of preferences, concerns, and predicaments of the different individuals within the society? How can we find any rational basis for making such aggregative judgements as "the society prefers this to that" or "the society should choose this over that" or "this is socially right"? [30, pp. 128–129]

From the logical standpoint, it is evident that we are in the framework of a *relational deontic modal logic,* concerning alternative possible states of the *social* world that are *partially ordered* according to different rankings, depending on different physical, political, economic situations, but also on different value systems. I.e., different rankings of social states, which are satisfying different *maximality*—not "optimality", that is "maximally good for *all* the different context" and that as such is not *finitarily* computable—criteria of goodness for the different individuals and groups.[7] From a formal standpoint, this means that we are in the *Category Theory* (CT) framework of *computational topology* applied to ML [31].

Logically, we are indeed in the framework of the algebraic interpretation of *Kripke's relational modal logic*, based on topologies of *Non-wellfounded (NWF)*

[6] Indeed, Sen decided to dedicate his Nobel Lecture to illustrate this novel discipline, conscious of its relevance for the future of the social, economic, and political sciences. Indeed, it applies the axiomatic method also to them, and to their mathematical and experimental statistical tools, so to make them properly "sciences" according to the modern Galilean sense of the term. In few words, SCT is the branch of formal philosophy concerning the social world [29].

[7] In this sense, the implementation of Sen's theory requires Kripke's relational modal logic in its deontic interpretation, in which *local truths* are allowed, as far as coalgebraically modelled over topologies of *partially ordered* sets (see Sect. D1 in Appendix D).

sets [32], in which no *set total ordering* is admitted but different trees of *partially ordered sets* sharing the *same root*. This allows to define several FOL *local* semantics of modal *Boolean Algebras with Operators* (BAOs), each defined on a different *partition* of a given universe of possible world states (see [33, 34] for more details, and synthetically Sect. D1 in Appendix D).

Indeed, generally, Sen's SCT distinguishes among the different social theories of justice in economy and in politics, in terms of the *basal space* of the main variables with which each theory is concerned, and in terms of the *aggregation principle* of such variables characterizing each theory, and then discriminating between just/unjust states, on which the consequent social choices are justified.

For instance—to understand the utility of Sen's approach to SCT for formalizing different ethical and political theories in social science—in the *utilitarian* theories of justice typical of the liberal economy [1, pp. 139–140], the "basal space" consists "in the combination of the utilities of the different individuals, and nothing else—rights, freedoms, opportunities, equal treatments—is valued except for instrumental reasons". Consequently, the "aggregation principle", discriminating between just and unjust states in such theories is the simple "utility sum-total" for assessing the social state ("sum-ranking"), without considering other relevant factors, such as measures of "dispersion", or of "inequalities in accessing to opportunities", etc.

Sen's theory of the *comparative distributive* justice, on the contrary, substitutes the abstract and ineffective *Paretian equality principle* typical of the classical liberalism[8] with an *equity* or *fairness* principle that he borrowed from his main teachers: Aristotle, Adam Smith, and overall, John Rawls [35]. Sen's fairness theory, indeed, starts from the concrete evidence that groups and individuals *do not share* the same access to economic commodities and utilities, and *do not have* the same possibility of influencing the social choices. This depends not only on "manifest injustices" in the society, either on a national or international extension, but also on different ethical principles, and then on different evaluations of how the economic utilities and commodities are functional to "valuable and valued ways of living and behaving" or *functionings* in Sen's jargon, in view of the flourishing of the different personal *capabilities*.

Therefore, the "basal space" of Sen's theory of justice consists,

> in the set of combinations of functionings from which the person can choose any one combination. Thus, this "capability set" stands for *the actual freedom of choice* a person has over the alternative lives that he or she can lead. [14, p. 357]

The "aggregation principle" in Sen's distributive theory of justice is the Rawlsian famous fairness criterion of the *maximin* [36, p. 266]. That is, assigning more resources to the less advantaged individuals and groups in the society, to level the

[8] According to this *equality* principle whichever individual or group in the society have the same possibilities of accessing/enjoying the basic liberties, as well as the economical utilities and commodities, and then they have the same possibilities of influencing the social choices/assessments (see [37] for a systematic usage of this principle in SCT). It is evident the *unrealistic character* of this principle that is at the basis of the theory of *liberalism* and that is the formal root of its crisis in our liquid society (see [10] for a wider discussion)!

inequalities derived from "the natural lottery" theorized by Adam Smith, which blindly distributes talents, resources, and access to opportunities.

The main difference with Rawls' theory of "justice as fairness" [120], consists in the *comparative* character of Sen's theory. This ultimately depends on the fact that the maximin principle in Rawls' theory concerns the just *institutions* [116], while in Sen's theory the principle must work as an *aggregation principle of variables* in a SCT (see [10, 14] for more details). This means that, while in Rawls the maximin principle is intended according to the *Kantian normativism* (i.e., supposing the *absolute* character of *all* moral norms, for *all* possible contexts), and for which Rawls supposes a hypothetical "original position" in which an (ascetical) "veil of ignorance" is posed over all the *subjective differences* among humans and groups, so to consider all on the very same footing [36, 37], in Sen's SCT based on the interpersonal comparison of welfare states, we start from the relevance of the *subjective differences* [10]. These include not only the injustices and the inequalities in the resource distribution/access, but also the ethical and the cultural differences and preferences, the religious beliefs and even the personal tastes.

In this connection, one of the main results formally obtained by Sen in SCT is his demonstration that the inconsistencies derived from the usage of the Rawlsian maximin principle in SCT ultimately reduce themselves to the usage of the *Paretian axiom of equality* (see Note 8) in assessing the different welfare rankings in a SCT inspired by maximin fairness criteria. Sen demonstrated, indeed, that this axiom is the formal root of whichever "impossibility theorem" in SCT [36], the famous "Arrow's impossibility theorem" included [37], with its troubling consequences for the same notion of representative democracy, and that ignited a wide discussion in SCT literature.[9]

In Sen's theory, indeed, the Paretian axiom is substituted by the *axiom of extended identity* among individual positions, by which the maximin fairness criterion becomes an effective principle of variable aggregation, based on the comparison between different individual positions in different situations, so avoiding the systematic risk that a uniform application of the maximin becomes a source of effective injustice and of economical regression.[10]

This axiom wants to be a formal version of Adam Smith's *extended sympathy* principle, in the sense of "placing oneself in the position of another", extended to a society of individuals (see [14, pp. 210–220]). Sen gave a formal version of this principle in terms of *an axiom of extended identity* among n distinct welfare

[9] Roughly speaking, *Arrow's Impossibility Theorem* demonstrates formally in SCT that, given the Paretian axiom of equality, no shared ranking of welfare states and then of social choices is possible without some form of *dictatorship*. In this way, the early fame of the young Sen depends on the publication of his formal demonstration that Arrow's impossibility theorem is effectively a "a theorem of impossibility of a Paretian liberal" [36].

[10] Recently, because of the economical emergencies related with Covid-19 pandemic, all governments in the world applied uniformly the maximin principle for restoring at least partially the incomes of individuals and companies. However, it is evident to all that without applying comparative discriminative criteria (variable aggregations) among the different positions, this type of supports is not only economically unsustainable in the long-term, but also source of injustices.

rankings and relative positions of persons and groups in SCT. Indeed, as far as the SCT in Sen's quantitative approach allows *a comparable grading* of "gains" and "loss"[11] of commodities and utilities for different persons and groups in different social positions, the axiom of "extended identity" allows us to use the maximin principle on a *relative* and not absolute basis as *an aggregation principle of different welfare rankings*, for *consistently assessing* fair social rankings of welfare states.

That is, whichever ranking is only a *partial ordering* of social welfare states, because no *total ordering* (complete ranking) might ever exist, like in Rawls' normativism, or like in a SCT applying the Pareto unanimity axiom such as Arrow's social welfare functions. On the contrary, Sen's application of the maximin criterion gives us a suitable quantitative parametrization of the welfare aggregates involved.

For this aim, Sen extended the quantitative *comparative justice grading* of Patrick Suppes to n individuals (see Note 11) because it allows to include coherently in the model both an *utilitarian* and a *maximin* criterion for variable aggregation overcoming the limits of both.

> If (the social state) x is more just than (the social state) y in the sense of Suppes (with the extended identity axiom imposed), then x must have a larger welfare aggregate than y (*utilitarian relation*) and also the worst-off individual at x must be at least as well-off as any individual at y (*maximin relation*). [14, p. 208]

The axiom in its formal version within SCT is illustrated in the Appendix C of this paper. Anyway, this conclusion of Sen is particularly significant for our aims.

> The capability approach is entirely consistent with reliance on partial rankings and limited agreements. The main task is to get the weights – or ranges of weights – appropriate for the comparative judgements that can be reached through reasoning, and if the result is a partial ranking, then we can make precisely those judgements that a partial ranking allows. ([14, p. 369])

Where, of course, Sen's "reasoning" in ME is not only the human one but also the "AI reasoning" of ML models of SCT, in which the *degrees of freedom*[12] of the probability distribution of utilities and commodities among the different groups of individuals satisfy a maximin fairness criterion of variable aggregation in the sense just explained by Sen.

Indeed (see Sect. B7 in Appendix B, Sect. D2 in Appendix D, and [34]), it is possible to use the *Doubling of the Degrees of Freedom* (DDF) principle characterizing a "deep-belief NN" [40] in its QFT computational interpretation for implementing Sen's fairness theory based on the maximin principle in an unsupervised

[11] Effectively, as explained at length in the Chap. 9 (pp. 203–209), and in a formalized way in the Chap. 9* (pp. 210–218) of [14], Sen is here referring to a fundamental contribution of Patrick Suppes to SCT, in a paper published in 1966 [114], where he developed formally a "social decision function" based on the principle of *a grading of different level of justice*, on an *interpersonal* and then *equitable* basis, even though applied only to a two-individuals case. I.e., Suppes' rule is not properly a *social* choice function.

[12] We recall here briefly that the "degrees of freedom", as a result of a suitable variable aggregation, define in statistics the dimensions of the "probability space", within which a given probability distribution can variate (see Sect. D2 in Appendix D for more details).

ML algorithm to eliminate statistical biases in data. The DDF principle gives, indeed, an immediate computational effectiveness to the related Sen's "extended identity axiom" (see Appendix C) between the "basal spaces" (i.e., in Sen's terms "the sets of combinations of functionings from which persons can freely choose any one combination") of different disadvantaged/advantaged groups *balanced* into one only "fair" social state space of opportunity access to favorable social states. Indeed, also the DDF principle is in physics a *balancing principle* between two spaces of probability distributions representing a system and its environment, granting by a suitable "variable aggregation" in the resulting merged space, a sort of "fair distribution" of the resources (free energy) among all the components of such a doubled system (see Sect. D2 in Appendix D for a physical explanation of the DDF principle in the formalism of the computational QFT).

On the other hand, it is easy in the light of the discussion developed in Appendix D to guess that a physical counterpart of the maximin principle in economy for a fair distribution of resources is in the "fair distribution of energy" among the components of a complex dissipative system balanced with its thermal bath in physics. Therefore, Sen's *extended identity axiom* between different subjective basal spaces in SCT $(R_i, \tilde{R}_j)$, has in the DDF principle $(A, \tilde{A})$ of dissipative QFT its natural implementation as the basis of an unsupervised quantum ML algorithm inspired to the dissipative QFT underlying brain network dynamics (see Appendix D and [34] for more details). Or—if we prefer to use the "first-person" jargon of the intentional language for expressing Smith's "extended sympathy" principle (see Sect. B1 in Appendix B and the "cognitive triangle" of Fig. 4)—, only by mirroring "myself" in "you" so to be each "the double" of the other, we can constitute a sympathetic "we" (see [40]).

4 Conclusions: A Relational Ethics for Ethically Accountable AI Systems

To conclude, in our paper we discussed which are the logical and ontological conditions that AI autonomous systems must satisfy to be consistently considered as *artificial moral agents,* i.e., as *subjects of moral agency,* that is, of decisions with an ethical/legal relevance in the realm of ME. And not only as simple *objects,* i.e., as tools designed and used by humans—and then as simple extensions of the human intelligence and the human moral agency—, to which no autonomous moral agency can be attributed.

In our analysis, we started therefore from the basic notions of the *Theory of Active-Control Systems,* i.e., of *Cybernetics* as "the theory of communication and control in animals and machines"—according to Wiener's early definition. This reference is fundamental for recalling that moral agency can be attributed to humans and machines only and only if the "active control" on their actions concerns the ultimate supervising level of the *goals* of actions and of their "heterarchy", as McCulloch and

Pitts first emphasized in their pioneering work on ANNs [40]. That is, we can speak of moral agency in humans and machines if and only if their active control concerns the "targets" to be satisfied by their actions.

In this perspective, the ever-stricter interaction humans-machines in our Communication Age, must be interpreted as the interaction between *conscious* and *unconscious* communication/moral agents, respectively. In this framework, we emphasized that the notion of "distributed responsibility" between humans and machines in contemporary AI ethics discussions should be enriched by a further distinction derived from NE. This distinction is fundamental especially for autonomous AI systems endowed with "deep" ML algorithms and then with an unavoidable "opacity" in their decision processes (e.g., the "self-driving" cars). Indeed, the *slow* moral conscious responsibility of a skilled (trained) human driver concerns the moral/legal relevance of her/his driving actions. However, this does not mean being conscious of the *fast automatic* responsiveness of the modules of the sensory-motor cortices of her/his brain to the environment constraints (path curves, obstacles, etc.), ultimately constituting the driver skill. In other words, also in human the decision process by which our brains produce their fast adaptative response to the environment constraints are unavoidably *opaque* like in AI systems endowed with deep ML algorithms.

Therefore, instead of speaking about "distributed responsibility" between humans and machines, it should be better to speak about their *distributed accountability* to moral/legal constraints. More precisely, a distributed accountability between the slow responsibility of conscious moral agents and the fast responsiveness of unconscious moral agents to the ethical/legal constraints on their actions from the shared social environment.

Hence, the ethical/legal accountability of AI autonomous systems in ME and the connected "right to transparency" about their decisions that also AI systems must satisfy with respect to the society, does not concern directly the unavoidable "opacity" that the AI systems endowed with ML models share with humans in their decision processes. As we discussed in this paper, in the light of the "imitation game" of the Turing test from which the same AI research program originates, the ethical transparency in AI systems must concern what is *before* and *after* the opaque decision process in humans and machines.

Hence, the main contribution of the present paper to the ME discussion, is that the conditions that the AI autonomous systems must satisfy to fulfill *the right to transparency* that the society must pretend from them when their decisions have an ethical/legal relevance, are essentially two. The same two conditions that therefore AI systems must satisfy for consistently attributing them the ontological status of *unconscious artificial moral agents* in ME.

1. *The presence of explicit ethical/legal constraints on their FOL decisions on the individual actions to execute*. They can be implemented, either in the form of suitable deontic algorithms in the inferential-trees characterizing the program of a *symbolic* AI system; or in the form of suitable ethical/legal clauses to be satisfied in the ML algorithms of a *pre-symbolic* AI system. Indeed, both the *supervised learning process*, and the *deontic obligatoriness* as distinct from the *alethic* (logic,

causal) *necessity* in modal logic, ultimately consist into the calculation of an *optimization function* with respect to a given target/label (i.e., the minimization of some "cost function").

2. *The presence of an explicit "deontic reasoner" performing a deontic HOL automatic assessment* over the effective ethical/legal compliance of the FOL decisions taken by the AI system. That is, an *automatic* explicit metalogical deontic analysis over the decisions taken by the AI system endowed with some ethical competence, before that these decisions are transformed into *actions* over the social environment.

In our discussion, we emphasized that, while there exists already a wide literature about different implementations of ethical/legal clauses in the optimization function in which any supervised ML algorithm ultimately consists, only recently the AI research started to propose suitable solutions to satisfy the second condition we outlined. However, this second condition is fundamental for granting the ethical/legal accountability of AI autonomous systems for their decisions/actions, before all because also for humans it is the same.

Indeed, this sort of moral "self-auditing" is what characterizes our *moral reasoning*, as distinguished because following the *moral judgement* we made over the "right action to execute" before executing it. Effectively, the ethical/legal accountability for our actions depends mainly on this explicit (conscious) metalogical assessment (moral reasoning) about the morality of the judgement we performed on the right action to execute. It is therefore necessary to satisfy a similar condition also for granting the ethical/legal accountability for the decisions/actions of AI autonomous systems, so that we can consistently attribute to them the ontological notion of *artificial moral agents.*

Finally, particular attention we dedicated to the issue of the statistical bias toward given disadvantaged groups in the society which are present in the statistical samples on which the training of a supervised ML model is performed. This determines unwanted but effective "algorithmic injustices (unfairness)" in the decisions of the AI systems trained on these biased data. This issue is largely discussed in literature because strictly related to the precedent one of the "opacity" of the net training process. Generally, it can be solved by inserting *fairness* ethical criteria in *the statistical data pre-processing* of the training set of a supervised ML model. Effectively, such a pre-processing of the training set by an unsupervised ML algorithm for a suitable variable aggregation is anyway necessary for granting to the net a fast and reliable convergence to the desired results (see Sect. B7 in Appendix B and [77]). In our case, it must be performed through a suitable unsupervised ML algorithm satisfying a "fair" variable aggregation criterion, in the context of a *relational approach* to ME.

Therefore, in this context of a *relational ethics approach to deontic logic*, we devoted particular attention to the possibility of implementing Sen's theory of distributive justice as fairness. This is based on the ethical maximin principle used as a variable aggregation principle in the statistical data management, implemented as a particular architecture of QFT quantum computing for unsupervised ML. Indeed,

as explained at length in the Appendix D, the DDF principle of variable aggregation that is typical of the "dissipative" QFT modeling of quantum computing can be directly used as an unsupervised ML algorithm for defining a "socially fair" probability space, in which the statistical distribution function outputted by the supervised ML algorithm of the AI system can be defined. This modeling, indeed, *from a formal standpoint*, is particularly suitable for a relational ethics approach to ML. Indeed, on the one hand, it is directly inspired to brain unsupervised learning modeled using the dissipative QFT as fundamental physics of the brain dynamics. On the other hand, it is compliant with an algebraic modeling of deontic logic using Kripke's relational semantics of modal logics.

Appendix A: The Notion of "Active Control" in Biological and Artificial Systems and Its Relevance for ME

A1. The Graded Notion of Active Control in Biological and Artificial Systems

In his *Cybernetics* book [15], N. Wiener introduces the famous ontological distinction between "mechanical systems" that are capable only of a *passive control* on their own behavior (i.e., satisfying the "Third Action-Reaction Principle" of Newtonian Mechanics, and then ultimately a stability condition at equilibrium in classical and statistical mechanics) and the "cybernetical systems" that are capable of an *active-control by feedback* on their own behavior.[13] Where the notion of "feedback" consists in the fact that only a part of the output y (and then a "*physical* signal") is backpropagated toward a *controller C* able to modulate the input x, for (recursively) minimizing some measurable *distance* Δ between the *output value* y_e and *a target value* y_t (see Fig. 2).

As discussed elsewhere [38], this basic *triadic structure* (input-controller-output) of any active-control system—where the controller plays the *semiotic* role of Peirce's *interpretant*—is able to transform a "*physical* signal" into a "*communication* signal" or a *sign* (i.e., "something being for something else"). That is, a physical signal carrying on some "information" (i.e., where the "energy" measure is distinguished from the "information" measure). This makes possible to justify—in C. Shannon's statistical theory of communication—the distinction and the strict relationship (because both related to some minimization of the free-energy function, even though

[13] Following a famous exemplification of the physicist Victor F. Weisskopf [117] in the atomic and molecular physics, the physics of the atom nuclei controls the structure of the electron distribution according to different levels of energy (the electron "orbitals" according to the semi-classic Bohr's "planetary" representation of the atoms) around the atomic nuclei, but there is no feedback from electrons over the nucleons (protons and neutrons) in the atomic nuclei. Thus, following Weisskopf suggestion, atoms and molecules, despite their complex structures, are not active control systems, differently from also the more elementary biological system like a cell or unicellular organisms.

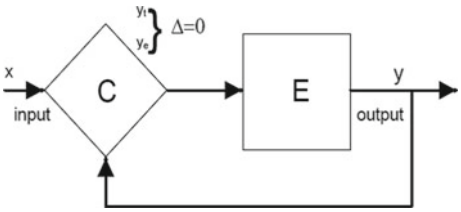

Fig. 2 Elementary scheme of an *active-control* system

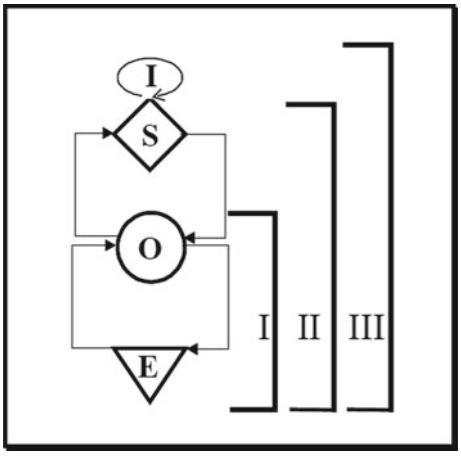

Fig. 3 Schematic representation of the three main levels on which an active-control can be exerted in biological and artificial systems

not necessarily at equilibrium) between the *physical* entropy *S* and the *information* entropy *H*, both sharing the same statistical definition. In this way, Wiener was right in his visionary approach aimed at giving by the notion of *active control* and its sophisticated mathematical apparatus, for the first time in the history of modern science, a strong common mathematical basis, both to the biological sciences and to the artificial sciences.

This common mathematical basis of the notion of active control in biological and artificial systems—following Müller's suggestion—can help us also in clarifying the notion of "graded and relative" autonomy of AI systems/robot with respect to the human control. Indeed, always referring to the basic notions of cybernetics, it is well-known that the active-control can be exerted both in artificial and biological systems (humans included) at three main levels:

1. The active-control over the *execution/not-execution* of some operation by the *effector* sub-system *E* (see Fig. 3) in the more elementary active-control systems (think, for example, at a simple thermostatic switch).
2. The active-control over the *organization level O* (see Fig. 3) of the complex response of a system endowed with several types of sensors for taking into account different parameters. Think, for example, at the "smart thermostats" with several sensors of the modern refrigerators. This is typical in nature of the biological systems as *self-organizing systems,* because endowed with *non-linear*

self-regulation processes at different degrees of complexity. This makes them able of adapting themselves to a varying environment, so to be stable in *far-from-equilibrium* conditions (think, for instance, at the biological *homeostasis* [39]).

3. The active-control over an *heterarchy* of the goals (targets) to be fulfilled, which *supervise* the behavior of any active-control system. It is evident that we speak about *autonomous* systems both in the biological and in the artificial systems, when the active control concerns this ultimate *supervising* level *S* in Fig. 3). E.g., in the self-controlled behavior of the human free-agency and/or in AI autonomous systems. Think, for instance, at the typical ethical conflict in self-driving cars between the constraints of the safety of the car passengers, and of the safety of a pedestrian crossing suddenly with a red-light a narrow street with high walls on both sides.

This active control at its ultimate level is generally implemented in *two fundamental ways* in AI systems, even though in many practical applications there exists an effective hybridization of the two approaches (see Appendix B):

1. In *symbolic AI systems*—the so-called "expert systems" because simulating algorithmically the expertise of humans in some specific field of data management (see Sect. 2.2)—by inserting the ethical constraints in the form of *deontic logic algorithms* in the explicit inferential tree, on which the system decisions of this class of AI systems is based without any "opacity" (see [19] for an updated bibliography about this type of approach).
2. In *pre-symbolic AI systems,* that is, the AI systems endowed with ML models due to the extremely large amount of data to be managed (often with millions of items and billions of parameters ("big data")) that excludes in principle any human expertise. In this case, the ethical constraints are inserted as ethical conditions (deontic logic "AND" clauses) to be satisfied in the optimization process (minimization of the error), to which any multilayered supervised ML algorithm ultimately reduces itself. This implies the ineludible "opacity" in the AI system decision process (see [20] for an introduction and [40] for deeper considerations on these topics).

To conclude, whichever system, either natural or artificial, able to implement an *active control over the ultimate level* of the goals of its behavior is effectively able to perform *deontic modal logic calculations,* which connotes it formally as a *natural/artificial moral agent.*

A2. The Deontic Modal Logic as a Formal Justification of the "Hume Law"

In the axiomatic logic framework, *modal logic*—that is, the logic of the different senses of *necessity/possibility* in philosophical logic—ultimately consists in adding

some *modal axioms* to the usual axioms of the propositional calculus. These axioms rule the consistent usage of the *necessity* $\Box$/*possibility* $\diamond$ operators in the modal propositional calculus. This means that the modal logic is a *two-valued* (1/0) propositional logic in which the true/false evaluation function of complex propositions cannot be reduced to the usage of the *truth-tables* of logical connectives ("NOT", "AND", "OR", "IF…THEN") among elementary (subject-predicate) propositions like in the usual propositional calculus [41]. Indeed, the truth evaluation function in modal semantics depends on *different truth criteria*, according to the main different (alethic, epistemic, deontic) interpretations of the modal operators in different linguistic contexts/usages [42].

In this axiomatic framework, it is therefore possible to satisfy formally in modal logic terms the so-called "Hume Law". That is, the distinction between the "necessity" in the descriptive statements of the physical/metaphysical discourse, and the "obligation" in the normative statements of the moral/legal discourse.[14] That is, in terms of the modal logic distinction between *alethic* (physical, causal) *necessity/possibility* ($\Box/\diamond\Diamond\bigcirc/\Diamond$) operators, and *deontic* (moral, legal) *obligation/permission* (**O/P**) operators.

Particularly, the so-called *value-based deontic logic* interprets the *reflexive* modal relation of alethic logic based on the modal axiom $\mathbf{T}(p) := \Box p \rightarrow p$, that is, "if p is true in all possible worlds, then it is true in the actual one", in terms of satisfaction of an *optimality axiological condition.* That is, p is "maximally good" for a given moral agent x in situation: $\mathbf{Op}(x, p)$. Then: $\mathbf{O} := (\mathbf{Op}(x, p) \wedge x_a \wedge x_{ni}) \rightarrow p$.

For instance, in alethic contexts it is sufficient that, if a physical law p holds in all physical contexts, then it holds also in the actual one. E.g., in the case of Galilean Law, "if it is necessary that all heavy bodies fall, then they fall also in the actual world". On the contrary, this is not true in deontic context, for instance in the case of the moral/legal norm of tax payment. That is, it is not true that "if it is obligatory that all people pay taxes, then all people pay taxes in the real world".

In order that the deontic obligatoriness of a moral/legal norm becomes effective in the real *social* world, it is necessary that this norm be related with a value to be pursued (something that is good or "optimal") for a given moral agent *x*. x must satisfy the double condition (clauses) of accepting it, x_a, and the freedom situation of having no impediment in effectively pursuing this goal, x_{ni}, by a suitable "good" action. Of course, in the case of pluralistic societies like ours, instead of a HOL condition of ethical optimality **Op** (= maximally good for all social-world states), it is sufficient a FOL condition of ethical maximality **Max** for given partitions (or disjoint unions or coproducts) of the social-world states. This brings us to a *relation ethics* based on Kripke's modal relational semantics (see also Sect. D1 in Appendix D).

[14] Historically, indeed, the Hume Law distinction resulted to be so impressive in the Modern philosophical debate, because of the abandon of the modal logic distinctions—well known and largely discussed in the Scholastic philosophy—in modern logic and philosophy from the XV cent on. Practically, till to the beginning of the XX cent., when C. I. Lewis proposed his axiomatic formalization of the modal logic [121], reinserting it in the modern philosophical and scientific debate.

Appendix B: From the Symbolic to the Pre-symbolic Approach in the AI Research Program

B1. The Origins of AI Research Program and of Cognitive Sciences

For a further clarification of what we intend when we affirmed that our solution of the opacity issue in AI autonomous system fully satisfies the "imitation game" of the Turing test, on which the AI research program is based, it is convenient to shortly review the same origins of AI program. The famous "Dartmouth University Conference" of 1956 started officially the *AI research program,* based on the *Turing test* (1950) and its "imitation game" [16]. It is impressive how the statement by which John Mc Carthy, Marvin Minsky, Nathanial Rochester, and Claude Shannon proposed to the Rockfeller Foundation to support this Summer Workshop at the Dartmouth College effectively anticipates the research program of AI developed during the following decades.

> We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. [17]

The proposal goes on to discuss digital computers, natural language processing, neural networks, theory of computation, abstraction and creativity, all research fields of AI based on the so-called "AI-dogma", as Douglas Hofstadter named it [43]. Namely, if a *Universal Turing Machine* (UTM)—effectively a "general purpose" programmable computer—can imitate successfully a human intelligent task, there *must* exist some essential "isomorphism" between the program running in the computer, and the program running in the brain. From this principle, the re-interpretation of the classical "mind–body" relationship in terms of the "software-hardware" relationship, and then the metaphor of mind as "software" of the brain "wetware", historically derives [44]. From this, in the 60's of the last century, the *cognitive science* research program as the "new" science of mind arises [45]. This is characterized by a non-reductionist approach, with respect to the physicalist approach of the so-called "central-state theory" to the mind science, proposed by Herbert Feigl at the end of 50s [46].

Feigl, indeed, was one of the youngest members of the *Wiener Kreis*, cradle of the "neo-positivistic movement" in Europe at the beginning of XX cent. After that he moved to US, founded at the University of Minneapolis the "Minnesota Center for Philosophy of Science". He also became the editor-in-chief of the prestigious collection of the "Minnesota Studies of Philosophy of Science" that outlived the death of his founder (1988) till today. In this collection, the results of the research

of the Center, and of the movement of the so-called "logical empiricism" originated from the Center activities, were published for several decades.[15]

Now, in the II Volume of the collection, dedicated to the mind–body problem, there were two fundamental contributions. The first one was the already quoted Feigl's article. The other one was by Wilfrid Sellars and was dedicated to the logical analysis of the relationship between "the intentional and the mental" [47]. In it he rightly emphasized that the "first-person (singular/plural) language" (i.e., the so-called *I/we talk*) expressing the "intentional (with "t") states of mind" of individual/collective cognitive *subjects*, supposes an "intensional" (with "s") modal logic". This makes *logically inconsistent* any materialistic attempt of identifying by a logical equivalence, an intentional state of mind, with an observed state of brain. The observational language of science, supposes, indeed, the standard "extensional logic"[16] of the mathematical pure and applied sciences. In a word, the *first-person* "I/we-talk" of the intentionality cannot be reduced systematically to the *third-person* "O-talk" of the observational language of the neurophysiological sciences, in their searching for the neural correlates of subjective mind states.

Sellars' distinction between the "I/We-talk" of the mentalistic language expressing the intentional conscious states, and the "O-talk" of the observational language of the neurophysiological inquiry influenced systematically the further philosophical reflections about the relationships between the intentional mind and the brain. From another philosophical standpoint, Willard V.O. Quine synthesized the issue in the following statement. We passed from Descartes' "irreducible duality of substances" about the mind–body to the "irreducible duality of languages and their logics". Even though both languages are sharing the same extra-linguistic referent: the physical states/operations of the brain [48, pp. 132–134].

---

[15] It is worth to be recalled, that Karl R. Popper in his *Intellectual* Autobiography, defined himself as "the killer of the Neo-Positivism", identifying the date of such a murder with the (temporary) stopping of the publication of the Minnesota Studies collection. Unfortunately for him, the collection (not the Neo-Positivism) outlived not only Feigl's death, but also Popper's death (1994).

[16] I.e., where the *extensionality axiom* holds between classes **A, B** holds: $((\mathbf{A} \leftrightarrow \mathbf{B}) \Rightarrow (\mathbf{A} = \mathbf{B}))$. That is, where the predicative meaning reduces itself to the predicate extension and then to the set-theoretic logical *membership*, so that two predicates (e.g., "being water" and "being H_2O") with the same extension (defined on two equivalent classes of objects) must be considered as *identical,* and then can be substituted each other, without changing the meaning of the predicative sentence [118]. In *intensional logics*, the extensionality axiom of mathematical logic does not hold, because what the individual/collective intentional subjects *intend* with a given predicative sentence is fundamental [119]. Formally, the different *intensional logics* (mainly, the *ontic*, *epistemic* and *deontic* logics) are different *semantic interpretations* of the common underlying syntax of the *axiomatic modal calculus.* This is obtained from the classic propositional calculus, by adding some modal axioms, ruling the usage of the "necessity" $\Box$, "possibility" $\Diamond$ modal operators [41, 42]. In other words, all intensional logics—constituting the core of the so-called *philosophical* logic (i.e., the logics of the ontological, epistemological, ethical disciplines where the reference to the human conscious subject(s) is essential), as distinguished from the *mathematical* logic of the scientific disciplines—are not "truth-functional" based on the usage of the "truth tables" of the logical connectives like in the mathematical propositional logic. Each intensional semantics is indeed characterized by a different *truth criterion,* i.e., by a different interpretation of the modal operators, through which different *intensional* logics are distinguished (see Sect. A2).

A similar ontological position was shared also by Feigl in [49], where he proposed his physicalist interpretation of the "central-state theory" of the mind–body relationship based on Sellars' irreducibility of the intensional to the extensional logics. Feigl suggested that an appropriate mind's science must be based on a triangulation among:

1. The *I-talk* of the intentional mentalistic language.
2. The O_1*-talk* of the observational language of neurosciences that he denoted as "physical$_1$".
3. The O_2*-talk* of the observational language of behavioral sciences that he denoted as "physical$_2$".

Now, what characterizes Feigl's central-state theory is the relationship between the two observational languages denoted as "physical$_1$" and "physical$_2$". Feigl assimilated them to the relationship in thermodynamics between, respectively, the "microstates" of the particle motions, and the correspondent "macroscopic" thermodynamic statistical variables (temperature, pressure, and volume). These have their proper explanation at level of the microstate dynamics, which in our case is the microstate of the brain dynamics. However, what links Feigl's theory to the early AI-research program and to the development of cognitive sciences and neurosciences is the observation that, both the *physical entropy S* in Boltzmann's statistical thermodynamics for closed systems, and the *information entropy H* of Shannon's *mathematical communication theory* applied to TM computations, share ultimately the *linear* character of the dynamics involved [49].

This supposition identifying *energy* and *information*,[17] evidently, no longer applies when, starting from the 70s of the last century, the *strong non-linear character* of the thermodynamic processes experimentally emerged. They indeed characterize all the "open" or *dissipative systems*, such as all the biological systems and mainly the natural brains are. On this basis, Walter Freeman [50], John Searle [51, 52], Hubert Dreyfus [53], all working at University of California in Berkeley, strongly criticized the early AI approach in cognitive neuroscience, necessarily based on the UTM linear computations. In a naïve but substantially correct way, Searle by his famous "Chinese room" metaphor, as opposed to the "room" of the Turing test [51], stated that a UTM is not a model of the human brain because brains calculation are based on the *intensional logic*—the logic of the psychological intentionality, as Sellars taught us—and not of the *extensional* mathematical logic of a TM (see Note 16).

Therefore, following Howard Gardner's historical reconstruction of the early development of *cognitive science* [45], what characterizes the "cognitive revolution" in the mind science is the complete substitution to Feigl's "physical$_2$" observational language, with the *computational language* of the information processing in the brain dynamics. In other terms, the updated triangulation of the modern cognitive neuroscience, as well as of the AI systems in *Theoretical Computer Science* (TCS), is among [54]:

[17] It is remarkable that A. Einstein stressed that in fully deterministic systems the energy-information distinction has no sense at all.

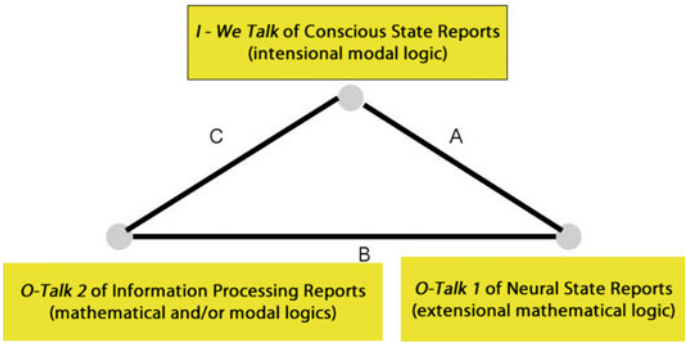

Fig. 4 Scheme of the triangulation of the cognitive neurosciences

1. The *I/We-talk* of the subjective intentional state reports in "singular/plural first person". They are formalized in the "intensional logics" like as many ("ontic", "epistemic", "deontic") interpretations of the *modal calculus.*
2. The *O-talk*$_1$ observational language of *neuroscience*, formalized in the *extensional logic* of the neuroscience mathematical models.
3. The *O-talk*$_2$ of the observational language of the *information processing* in the brain. They can be developed, either in terms of the mathematical calculus of the *extensional logic*, or in terms of the modal calculus of the *intensional logics.*

Of course, what is interesting for our aims (see Appendix D) is the possibility of implementing in AI systems the deontic algorithms of a *modal BAO* that has its proper foundation in the topological approach to TCS, based on the fundamental Marshal Stone's "Representation Theorem of Boolean Algebras" [55]. And then in the consequent development by Alfred Tarski and Bjarni Jónsson of a "Boolean Algebra with Operators" (BAO) that allowed the extension of the *operator algebras* formalism from physics to logic (see Sect. D2 in Appendix D and [54, 56–59] for further details). Particularly, in the framework of the *Category Theory* (CT) metalanguage, the *functorial dual equivalence* between the category of *coalgebras on Stone Spaces* **SCoalg,** and the category of *modal BAO* **MBAO** for the *Vietoris functor* $\mathcal{V}$, i.e., $\mathbf{SCoalg}(\mathcal{V}) \simeq \mathbf{MBAO}(\mathcal{V})^*$ has a particular relevance, because it allows an *algebraic interpretation* of Saul Kripke's *modal relational semantics* [60, 61], with a direct applicability in TCS [33, 62]. Moreover, there exists the possibility of an implementation of this functorial duality in a categorical modeling of quantum information processing in dissipative QFT systems, both in cognitive neurosciences [63, 64] and in TCS [54, 65]. This possibility is based on the identity of the topological properties between the Stone spaces in logic and the Banach spaces of the C*-algebras in quantum physics [66]. Particularly, this means that it is possible *a quantum implementation* of the so-called "deep-belief neural networks", as a particular model of unsupervised ML developed by Walter Freeman and his colleagues in AI systems [67] (see Sect. B7 in Appendix B). In our case, it can be directly applied to implement

in AI autonomous systems the deontic algorithms of a *relational ethics,* rigorously formalized in the framework of Kripke's modal relational logic, as we discuss in Appendix D.

B2. The Symbolic AI and the Functionalist Approach in Cognitive Sciences

For continuing our reconstruction of the AI research program, the interpretation of the information processing in the brain in terms of the UTM calculations, is what characterizes the early *functionalist* approach to cognitive sciences. This has its manifesto in the already quoted paper by H. Putnam—who successively changed completely his mind—its manifesto [44]. On the other hand, this approach has its development in the so-called *symbolic approach* based on UTM to AI [68–70]. Indeed, in this approach the decisions of an AI system are based on the *explicit inferential decision tree* implemented by the programmer, without any ML algorithm, and then without any "opacity" in the decision process of the system.

Furthermore, the pioneering work of Warren S. McCulloch and Walter H. Pitts during the 40's of the last century [71] demonstrated that a (net of) neuron(s), with a *linear activation function*[18] can in principle implement the four basic Boolean logic operations ("and", "or", "if.. then", "if and only if… then"), and then it is equivalent to a TM. In this way they extended the symbolic approach of AI, and then the functionalist approach to cognitive sciences, also to *artificial neural networks* (ANN).

Moreover, at the end of 40's, Donald Olding Hebb, based on neurophysiological evidence, defined the so-called associative *Hebbian learning rule* for the self-assembling of neuron circuits in the brains [72]. According to it, the recurrent simultaneous activation of neurons produces an increasing in the *synaptic statistical strength (weight)* among these neurons, following a *linear* rule. A rule that can be, therefore, synthesized into the slogan "neurons increase the probability of wiring together if they fire together" [73]. In 1954, B. G. Farley and Wesley A. Clark published a computational model of *self-organizing ANN* based on the Hebbian rule, in which arrays of artificial neurons (effectively, the cells of a *transitive probability matrix* (TPM)[19] simulating on a digital computer the net dynamics) are enriched by feedforward/feedback circuits, determining the *statistical weight of connection* w_{ij} between two neurons i, j. These works, therefore, inaugurated, inside the realm of cognitive sciences, the new discipline of *cognitive neurosciences* [74]. In this case, the

[18] The linear activation function (state transition map) for each neuron is given, indeed, by the algebraic summation of positive and negative input values, simulating the array of excitatory/inhibitory synapses of natural neurons. When the overall value overcomes a fixed threshold, the neuron is activated so to display a discrete 0/1 behavior.

[19] We recall that a TPM is matrix of *conditional* probabilities ruling the *transition* of the activation state (0/1) of each neuron (a cell of the matrix) in a way depending, according to a given statistical rule, on the activation state of the other connected neurons (the other cells of the matrix).

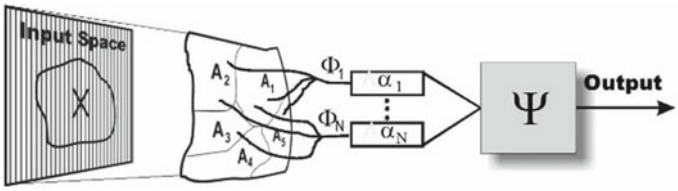

Fig. 5 Schematic representation of the linear perceptron parallel architecture, where a given pattern X is "designed" over the input space (or "retina") of the net. Each input neuron α_i calculates independently a different function Φ_i, whose supports are defined on a "filter" constituted by disjoint subsets A_i of the input space. The output neuron Ψ, therefore, calculates the simple linear summation of the results of the input neurons calculations

computing system of reference is a *probabilistic* TM, always with the supposition of the *linear* character of the statistical dynamics involved.

B3. The Pre-symbolic Approach to ANNs and the Linear Machine Learning

A further significant approach to early ANN architectures is the so-called *linear perceptron architecture* of Frank Rosenblatt [75], who during the 60s hoped to implement a *parallel computing architecture* in a net of linear neurons, for simulating the parallelism of the brain neural networking. The parallelism of the architecture depends on the fact that each neuron is calculating *independently* a different function defined on some disjoint subsets of the input set (i.e., mathematically a *filter* defined on the power set of the input set). This has evident advantages as to the standard *serial* computers in terms of computational velocity.[20] But overall, without the necessity of a *supervisor* (programmer) distributing the different computational tasks among the neurons (see Fig. 5).

In this way, the perceptron introduced the notion of (unsupervised) *machine learning* (ML) in the ANN research program. Indeed, the updating of the statistical weights (i.e., the probabilities of neuron activation) associated to the neuron connections, during the *training (learning) phase* of the net over a representative sample of the dataset, is based on a *linear* Hebbian-like rule.

However, in 1969 Marvin Minsky and Seymour Papert published at MIT a book with a strong criticism of Rosenblatt's linear perceptron [76]. Indeed, they demonstrated mathematically, and in a very elegant and convincing way, that for this type of parallel computational architecture conceived for the *pattern recognition* of one only class of objects, it is in principle impossible to calculate the logical "XOR"

[20] Indeed, if an algorithm is composed by n computational steps, while in a serial computer it is performed by one only processor in n cycles of calculus, in a parallel architecture it can be performed by n processors in one only calculus cycle.

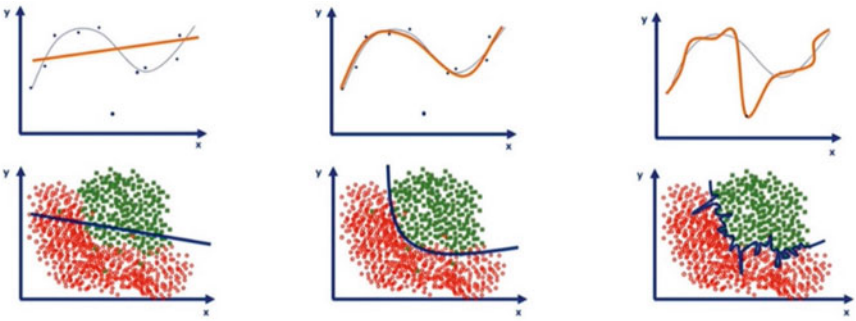

Fig. 6 Intuitive representation (from [77]) of the two errors of *under-fitting* (left) and *over-fitting* (right) at the end of the training phase of the ML algorithm of an ANN, for the discrimination between two classes (green and red) of objects. It is evident that the *best-fitting* (center) implementing statistically the logical XOR is given by a non-linear function, that a linear function (left) cannot implement in principle. On the contrary, the *over-fitting* is given by a function too depending on the training set and then with null generalization capacity, since it fits also with elements randomly distributed over the two classes in the training set (Finally, we recall that in any ML algorithm, to test the results of the training phase, the generalization capacity of the learned classification is tested on another representative sample of the dataset, distinct from the sampled set used for the training phase, in the so-called "testing phase" of the ML algorithm. Only after a successful test the ML ends, and the system is applied to perform its classification task on the whole dataset)

(or "exclusive or" (0110) corresponding to the negation of equivalence (1001) that is essential for training an ANN for executing *classification* tasks among objects belonging to different classes (see Fig. 6), which, on the contrary, it is simple to be calculated by a standard serial computer.

Moreover, a second criticism to the Rosenblatt perceptron was, if possible, even more radical. Indeed, the union of disjoint sets of the perceptron is a *proper filter* only and only if it is granted that at least one point of the "pattern" in the input space (corresponding to a given correlation order among the elements (points) of the input space) falls within one of the disjoint sets of the filter. Indeed, a *proper filter* is defined in set theory as a partially ordered set defined on the power set of a given set, *with the exclusion of the empty set*. Now, the only way for mathematically granting this "fitting" of a filter with any pattern designed in the input space is the existence of a "supervisor" seeing at the whole input space, readapting continuously the filter for matching different patterns designed into the input space. But in such a way the perceptron would lose its fundamental property with respect to a standard serial computer. That is, its pretension of being an effective parallel architecture of calculus. Now, Minsky's and Papert's criticism was so destructive because mathematically incontestable against the early linear approach to the ANN parallel computing that this book effectively blocked any attempt in the ANN direction till the 80s of the last century.

B4. The Pre-symbolic Approach to ANNs and the "Deep Learning" in AI Machine Learning

On the other hand, from the neurophysiological point of view, a lot of experimental evidence was produced during the 70s of the last century, emphasizing the *non-linear* and even the *chaotic* character—in the sense of the dynamic notion of *deterministic chaos*—of the natural NN information processing in the brain (see [78] for a synthesis). This determined the crisis of the functionalist approach to cognitive sciences from the standpoint of neurosciences, with the consequent refusal of the early "AI dogma" for which a linear (probabilistic or not) TM might be *always* a faithful model of the neural computational architectures in natural brains.

This determined a paradigm-shift in ANNs, denoted as the *connectionist approach* to ANNs, and then of the so-called *pre-symbolic* approach to ML in AI-systems, versus the early "symbolic" one like in Minsky's celebrated AI *frame theory* for expert systems [79], the ancestor of the actual *object-programming* techniques. The connectionist approach, indeed, is aimed at the statistical management of huge bases of data ("big-data") with higher-order inner correlations instead of the first order averages that can be calculated by a TPM endowed with a *linear* activation function [80].

Particularly, the so-called *backpropagation* (BP) machine learning algorithm [81] seemed to directly solve the core of Minsky's and Papert's criticism to Rosenblatt's linear perceptron, before all the ability of calculating the logical "XOR" function that is indispensable for classification tasks [14] (see Fig. 6). What characterizes the BP architecture as to the linear perceptron is indeed:

1. The presence, beside the only input and output layers of the original perceptron, of several *inner layers* of neurons, so to justify the notion of "deep-learning" in this type of ANN architecture.
2. The presence of a *non-linear* function (threshold) multiplying the activation function (i.e., the weighted input summation) of the deep neurons of a BP for a *sigmoid function* $\sigma(a) = \frac{1}{1+e^a}$, and/or by its close relative, the *hyperbolic tangent function,* tanh*,* instead of the stepwise 1/0 activation function of Rosenblatt's perceptron. This latter is effectively the *Heaviside function,* whose value is *zero* for negative arguments and *one* for positive arguments, and then making linear the neuron activation function (see Fig. 7).

In a word by using the sigmoid function or the hyperbolic tangent function in the activation function of the hidden neuron layers, the neuron output can be any real numerical value between 0 and 1, so to allow a *non-linear* characterization of the neuron statistical outputs, instead of the discrete 0/1 output of the linear stepwise activation function of the McCulloch and Rosenblatt neurons. This makes in principle the net able to perform more complex statistical computations, addressing to higher-order correlations (complex combinations of variables) in the input data set.

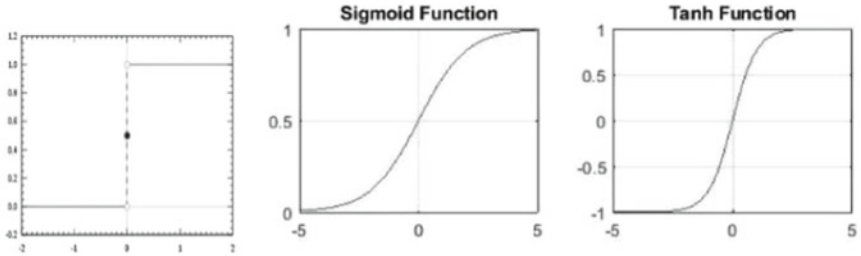

Fig. 7 Heaviside step function (left), acting effectively as a discrete 0/1 threshold, compared with sigmoid (center) and hyperbolic tangent (right) functions. It is evident that tanh is a 0-centered function with values between −1 and 1, then satisfying an anti-symmetric relation that, is fundamental for justifying the set-ordering in mathematical logic

B5. The Gradient Descent Algorithm in Supervised Machine Learning

Finally, the statistical output so obtained allows BP to use the *stochastic gradient descent* optimization algorithm, for the weight update of the hidden neurons during the learning phase of such an architecture, and from which its "back-propagation" name properly derives. Indeed, the *supervised* learning process of this multilayer non-linear ANN structure—developing an early suggestion by Paul Werbos [82]—consists into a stochastic optimization process of *error minimization.*

That is, the *supervised* "deep-learning" of the inner neurons of BP is modeled as the stochastic (random) searching for the *global minimum* of the "error-function potential" of the net weight dynamics. Where the error is substantially a Euclidean distance (effectively a measure of the "mean square error") between the "desired" probability distribution, and the "actual" probability distribution outputted by the net [81]. As schematically synthesized in the Fig. 8, at the end of each training cycle, the BP algorithm, estimates the error and "back-propagates" it for a re-adjustment of the hidden neuron weights to reduce the global error at the next step. And so on, recursively till the *global minimum* of the error function is reached.

Of course, the blind redefinition of the hidden weights among the different "neurons" each representing a different "variable" of the complex problem at issue, constitutes a big problem for the usage of AI systems as support for decisions implying social, moral and legal consequences. In fact, it makes *non-transparent* the data usage, as well as the motivations for which the system evaluates the legal/moral relevance of (i.e., "it weighs") each different components in relationship with the others as to the final decision. This blindness and opacity of the variable weighing are two of the main factors that ignited the actual fierce debate on AI ethics, as we emphasized in this paper.

Anyway, BP is not able in principle to answer properly the second main criticism of Minsky and Papert to perceptrons. The neurons of the inner layers, indeed, are connected "all to all" among themselves and overall, with the *input neurons.* In this

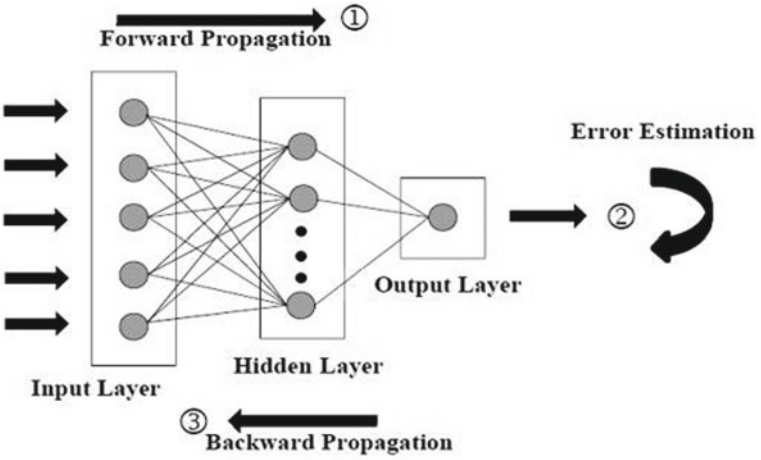

Fig. 8 Schematic representation of the BP "deep-learning" algorithm, according to which the net "back-propagates" the error from the desired output onto the weights of the hidden layer(s) in a random way, so to obtain recursively the global minimization of the error

sense, a connectionist neural architecture based on the BP algorithm is not properly an implementation of a parallel computational architecture.[21]

Indeed, it is mathematically true that the "smoothing" of the Heaviside by a sigmoid function, working in the BP algorithm as a fixed threshold on the weight activation function, is able in principle to calculate higher order correlations.[22] However, which they are—i.e., which are the disjoint arguments of the XOR—depends *critically* on the "slope" of the sigmoid that must be fixed in advance by the programmer. In fact, the learning process of BP is not on the neuron thresholds but on the weights, and, indeed, if we make varying also the thresholds (i.e., the connection topology among neurons) and not only the connection weights as it happens in natural neurons, the system dynamics becomes immediately *chaotic* [83]. On the contrary, as we demonstrated elsewhere [65, 84], the *dynamic definition of the sigmoid* by the *doubling of the degrees of freedom* (DDF) between the system and its environment is precisely one of the main characters of the dissipative QFT modelling of unsupervised learning both in natural and artificial NNs (see Appendix D). Not casually, indeed, it is demonstrated that what we observe *macroscopically*

[21] For this reason, Minsky refused to make any substantial correction to the Second Edition of his Perceptron book published in 1989 [76], vindicating—rightly from his theoretical point of view—that BP gave no substantial answer to his main criticism against the effective parallel computation capabilities of the ANN architectures.

[22] Indeed, in mathematical analysis, the Taylor series expansion of a tanh (and then of a sigmoid) function contains in principle all the correlation orders among the elements of a given set, and then in principle it can include whichever class defined on a given set of elements.

as a chaotic trajectory in the dynamics state (phase) space, is nothing but the trajectory among different phases of the *microscopic* quantum field dynamics that can be controlled by the DDF in dissipative QFT [85].

Then if the total connectivity and the related issue of the "sloping" of the sigmoid makes not properly "parallel" the BP machine learning, on the other hand, the total connectivity makes *extremely computationally heavy* the BP calculations. Indeed, the possible combinations grow factorially as $n!$ with the number n of the fully connected hidden neurons involved. This practical limitation determined therefore a second "latency" period of the ANN approach to ML in AI systems during the 90s of the last century till the beginning of ours. At that time, the large availability of non-expensive but computationally powerful (in the matric calculus) *graphic processor units* (GPUs) to be arranged in parallel architectures with many nodes is *one of the two factors* determining the actual explosion of the AI systems endowed with "deep learning" algorithms.

B6. The Deep Convolutional Neural Networks as the State of Art in Machine Learning

The *second more relevant factor* on which the actual explosion of AI systems endowed with "deep" ML algorithms was the publication of the paper by Geoffry E. Hinton and his Colleagues in 2012 on their model of *deep convolutional neural network* [86]. This model has successfully worked on the *Imagenet* database [87], containing millions of images (today, more than 14 millions). The model consisted of a convolutional neural network of nine layers of neurons, with 60 million parameters and 650,000 nodes that has been trained on about a million distinct examples of images taken from about a thousand classes.

Indeed, the convolutional neural networks (CNNs) are now the paradigm of reference in deep learning-based ML models. Or, in other words, CNNs are the main reason for the current success of deep learning in AI. The distinctive features of CNNs can be found on any good review paper of this type of connectionist network (see for instance [77], as one of the more recent and complete), of which there exist several models for different applications.

What universally characterizes CNNs compared to other connectionist networks are *two fundamental innovations*, making effectively more similar this ANN architecture to the networking of brain cortices:

1. The concept of *neuronal receptive field*. That is, each internal neuron of the *convolution layers* of the network sees only a *subset* of the respective input set from the previous layers just as in the original perceptron (see Fig. 5). Effectively the complete connectivity is only in the final classification layers of the network. In this way, CNNs can avoid systematically the problem of the total connectivity among the inner layers of the BP architecture.

2. The presence of *several inner layers of neurons of different types* (building blocks), according to the following general scheme:

   - *Convolution layers* for the progressive extraction of input features by *kernel operations* (filtering) that correspond to *different abstraction levels*. The output of each convolution operation is multiplied by some *nonlinear function* (generally the *Rectified Linear Unit* ($ReLu := f(x)_{ReLu} = \max(0, x)$), because of the linearity of the convolution operation itself.[23] Effectively in the CNNs a *discrete version* of the convolution operation between two functions $f[n]$ and $g[n]$ is implemented. Practically, each convolution layer contains a set of *convolutional kernels* (filters), "which is convoluted with the input image (with N-dimensional metrics) to generate a map of the emerging common characteristics of the input as output" [77, p. 523].
   - *Pooling layers,* each after a convolution layer for down-sampling the statistical output of each different convolution layer, to reduce the size of the output without losing significant information.
   - *Final classification layer*. It is the only one with neurons *totally connected* with those of the last convolution layer, and it is endowed with a function of *back-propagation of the error* on the convolution layers. In this way, the emergent property of a CNN during the training phase is that, both the classification level, and the convolutional levels of feature extraction *learn together*, so to make *evolving in time the same multi-layered filtering operations* (kernels). This is the *adaptive filtering* that characterizes the CNNs with respect to the perceptron, and that is the "secret" of the effectiveness of a CNN architecture (Fig. 9).

[23] The formal definition of the convolution in mathematical analysis is the operation between two functions which consists in integrating the product between the first value and the second one, shifted of a given magnitude. Formally, given two functions $f\ (t)$ and $g\ (t)$ defined on the reals $\mathbb{R}$, the following function is defined as convolution of f and g:

$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau$$

where $(t - \tau)$ is the shifting interval. Where the translation is *temporal*, the convolution practically corresponds to the *cross-correlation* operation. It explains, for example, in the visual system of the human brain, the role of the multiple crossings between the different nervous fibers from the cones and rods of the retina before converging into the optic nerve. The cross-correlation is effectively a measure of similarity between two signals, depending on the temporal translation applied to one of them. In this way, the visual system is able, for example, to extract the feature of the geometric shape of an object simply by cross-correlating between two different frequencies (colors) of the light radiation reflected by adjacent zones of the surface of the object and detected in succession by the cones and rods of the retina. And in fact, the feature of the edges thus extracted becomes the input of the inner neurons of the famous "area 17" of the visual cortex, which distinguish between horizontal, vertical, and oblique shapes of the visual object. For this fundamental discovery David H. Hubel and Torsten Wiesel earned the 1981 Nobel Prize in Physiology or Medicine.

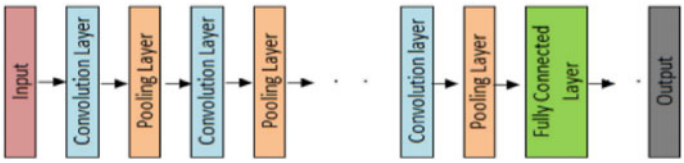

Fig. 9 Intuitive block diagram of a CNN architecture (from [77, p. 523])

Finally, *the error measure* for calculating recursively the gradient descent of the cost function in the supervised learning of a CNN architecture is generally the *cross-entropy function.* This is an alternative measure of logarithmic type to the more classical Euclidean loss function, i.e., the so-called "mean square error" measure, used in the perceptron and in the BP architectures.[24]

B7. The "Deep-Belief" Neural Networks, the Unsupervised Machine Learning and Its Relevance in Machine Ethics

A fundamental component of the training phase of CNN stressed by all the Authors (see for instance, [77, pp. 535–551]) is the *data preprocessing* of the training and test sets (see Note 21). This includes well established statistical techniques such as the *data normalization* and the *data augmentation*, with special care to the correct *parameter initialization* of the net.

A multilayer CNN model is indeed generally made up of millions or billions of parameters, so that the *proper initialization of weights* at the beginning of the supervised training process becomes essential to ensure, on the one hand, *the rapid convergence* of the model, and on the other hand *the accuracy of the result*.

Indeed, the simplest initialization technique of zeroing all weights is highly inefficient, so that generally the *random initialization* using casual matrices (i.e., using elements sampled from a Gaussian, or from uniform distributions, or from orthogonal distributions) is the normal choice. However, the best-performing initialization strategy of a supervised CNN relies on *a second unsupervised ANN* to give the CNN the initial weight values for its supervised training phase. Where we recall that "unsupervised learning" means a learning process in which the classes in which distributing the objects are not already defined or *labeled,* so that we speak of an "*unlabeled* learning process".

---

[24] Indeed, the cross-entropy measure generates the output within a probability distribution p, $y \in \mathbb{R}^N$, where p is the probability of each output category, y denotes the desired output, and N is the number of neurons in the output layer. The probability p of each output class i can therefore be obtained as $p_i = e^{a_i} / \sum_{k=1}^{N} e^{a_i}$, where e^{a_i} denotes the not normalized output from the previous layer of the network. Therefore, the measure of cross-entropy loss H can be defined as: $H(p, y) := -\sum_i y_i \log p_i$, where $i \in [1, N]$ [77, pp. 534–535].

Among the different models of unsupervised NN, the more efficient ones are the so-called *deep-belief NN* proposed by Robert Kozma, Marko Puljic, and Walter Freeman also because directly inspired to the unsupervised training of our brains modelled as dissipative systems [67]. Indeed, their unsupervised learning algorithm—giving the name to the model—is based on the principle of *a progressive clustering of significant variables* in the input dataset across the different layers of the network *to reduce the number of the degrees of freedom* (i.e., the dimensions of the probability space, in which a given probability distribution can variate) of the *output probability distribution* to those significantly *corresponding to the degrees of freedom of the input probability distribution.* A QFT version of the same approach to unsupervised learning—also because directly inspired by the same neurophysiological evidence—can be found in the *doubling of the degrees of freedom* (DDF) principle system-environment, discussed in the Appendix A.

For the aims of the present contribution, it is highly significant that suitable unsupervised models of ML are used also in ME for correcting the biases hidden in the statistical distributions, on which the training phase of autonomous AI systems is performed, with discriminatory "unfair" effects for the social minorities. The aim is, indeed, to make these models compliant with ethical criteria of *fairness* in ME, in the framework of a *relational approach* to a value-based deontic logic (see [23, 24]). The issue is well synthesized in the following quotation from a paper recently addressing the argument.

> Algorithmic assessment methods are used for predicting human outcomes in areas such as financial services, recruitment, crime and justice, and local government. This contributes, in theory, to a world with decreasing human biases. To achieve this, however, we need fair machine learning models that take biased datasets but output non-discriminatory decisions to people with differing protected attributes such as gender and marital status. Datasets can be biased because of, for example, sampling bias, subjective bias of individuals, and institutionalized biases. Uncontrolled bias in the data can translate into bias in machine learning models. [88, pp. 1, 2]

As we explain in the next Section and in Sect. D1 of the Appendix A, and as we discussed already in [34], we also suggest an approach to satisfy farness criteria in ML models that, as fair as, concerning data pre-processing, are not implemented as ethical constraints on the ML optimization procedure. However, differently from the precedent one that proposes a supervised ML procedure "tuned by-hand" for learning the fair model, we proposed an *unsupervised ML model* to data preprocessing for automatically correcting—i.e., dynamically, without any "fine-tuning" of the variables by the programmer—the biases in the training dataset, and so granting a "fair" ML model. Our approach, indeed, applies the DDF principle just introduced as an unsupervised ML strategy of data preprocessing to implement the core of formalized Amartya Sens's theory of fairness. Outstandingly, indeed, it uses mathematically the *maximin* principle (*max* of resources to *minus* advantaged) as a *fair variable aggregation principle* by which defining the degrees of freedom—i.e., the dimensions—of a "fair" social state space of equitable access to social/economic opportunities (favorable social states).

Indeed, just for this usage of the maximin as a variable aggregation principle, Sen can define in his mathematically formalized *Social Choice Theory* (SCT) as we see in Sect. 3.2, an *extended identity axiom* between the spaces of social states of disadvantaged and advantaged groups, *balanced* into one only "fair" social state space of opportunity access to favorable social states. Now, also the DDF principle is physically a *balancing principle* between two spaces of probability distributions representing a system and its environment, granting, by a suitable "variable aggregation" in the resulting merged space, a "fair distribution" of the resources (free energy) among all the components of such a doubled system (see Sect. D2 in Appendix A). Not casually, indeed the DDF characterizes the unsupervised learning process of our dissipative brains interacting with their physical-social environment, modeled in the Fundamental Physics framework of dissipative QFT [63, 64].

Appendix C: On Sen's Transformation Mapping of the Set of Individual States onto the Set of Social States

Synthesizing Sen's formal demonstration in [14, pp. 210–220], given the set X of individuals $i, j, \ldots$, and the set H of social states $x, y, \ldots$, and the Cartesian product $X \times H$ of all possible choices (that grows factorially with the number n of the individuals), to implement a fairness condition in the basal space of the individual $i \in D$, where D is a subset of disadvantaged individuals in X, it is sufficient to satisfy the following condition. That is, to extend the ranking R_i of welfare states among which i can exert her choice, to the *extended ranking* $\tilde{R}_i$ also including the ranking R_j of an individual $j \notin D$ (non-disadvantaged individual), because in this ranking there is also the state x that is better than y for i.

Formally, it means to impose the restriction of a one-to-one correspondence from the set of individuals H to H itself, such that $i = \rho(j)$, where ρ is *a transformation mapping* the (set of choices of) a person j onto (the set of choices of) a person i. The *restricted set* (partition, set disjoint union, or coproduct) of all this one-to-one-correspondences in X can be denoted as $T \subset X$ and justifies Suppes' assertion that "x is more just than y according to person i", $x J_i y$ in a restricted, and then *computable* way also when extended from two individuals—like in the Suppes' case—to n individuals.[25] That is, ρ is computable in terms of the restricted relation $x O_i y \leftrightarrow \exists \rho \in T : \left[\forall j : (x, j)\tilde{R}_i(y, \rho(j))\right]$. In other terms, given the transformation ρ, this justifies the person i in assessing that she prefers to be in the position x of *someone,* either j or i herself, than in the position of this same person in y. In

[25] Effectively, the number of the possible permutations in which a state x can be more just than y for n individuals grows as $n!$. Consider that, in terms of the DDF physical principle, Sen's restriction corresponds to the reduction of the number of the degrees of freedom of the two distributions to only those admitting the balancing principle of the minimization of the distance between a pair of states. That is, the minimization of the free-energy function (maximum entropy) between a pair of states (see Sect. A2 in Appendix A).

other terms, Sen's $x\,O_i\,y$ is the *necessary and sufficient condition* for justifying the consistency of the finitary computability of maximin principle of fairness in Sen's SCT, for whichever number n of individuals.

It is now possible for us to understand the statement of the *extended identity axiom* in SCT, as implementing the *relational ethics* principle of "extended sympathy" as necessary and sufficient condition for using the maximin as a variable aggregation principle in SCT on an *effectively* fair, equitable basis.

Axiom 1 (*Axiom of identity*) Each individual j in placing himself in the position of person i takes on the tastes and the preferences of i. That is,

$$\forall x, y \in X : \left[\forall i : \left\{(x, i)\tilde{R}_i^{\rho}(y, i) \leftrightarrow \forall j : (x, j)\tilde{R}_j^{\rho}(y, j)\right\}\right]$$

A stronger version of the Axiom 1 is the following "axiom of *complete* identity" identifying the rankings among all the persons belonging to the partition T, i.e.:

Axiom 2 (*Axiom of complete identity*). $\forall i, j : \tilde{R}_i^{\rho} = \tilde{R}_j^{\rho}$.

It is evident that for making formally consistent in abstract mathematics Sen's axioms of extended identity between the "basal spaces" of different social groups the topological notion of *equivalence by homotopy* is required.[26] Not casually, this notion of homotopic equivalence is at the basis of the emergent research field of the *computational topology* in TCS and then of the *topological data analysis,* recently applied fruitfully also to ML [31].

On the one hand (see Sect. D1 in Appendix A), this conclusion again emphasizes that the proper logic—in CT metalanguage—of Sen's relational ethics is within (the topological interpretation of) Kripke's modal relational semantics in terms of a coalgebra of NWF-sets defined on Stone spaces for a modal BAO semantics [33]. In it, partitions (set disjoint unions or coproducts) of admitted (social) states can be defined as "rooted-trees" of Kripke structures of possible states, so that it is possible to justify in this formalism a particular implementation of the homotopic equivalence in computational topology in terms of the notion of *bisimulation* (symbol: $\leftrightarrows$) between Kripke's structures/models (see [54, 89, pp. 53–55]).

Now, as we demonstrated elsewhere [54, 65], given that the properties of topological Stone spaces on which the algebra-subalgebras structure of a BAO semantics in logic, and the topological spaces on which the C*-subalgebras of Hilbert spaces in (quantum) physics are the same, it is possible to model a modal BAO semantics over

[26] Intuitively, in CT metalanguage, two different paths sharing the same endpoints x, y can be said "homotopically equivalent" if they can be *continuously deformed* into each other. More formally, given two topological spaces X and Y, a *homotopy equivalence* between X and Y is a pair of *continuous maps* $f : X \to Y$ and $g : Y \to X$ such that $g \circ f$ is homotopic to the *identity map* or reflexive morphism id_x and $f \circ g$ is homotopic to id_y. If such a pair exists, then X and Y are said to be *homotopy equivalent,* or of the same *homotopy type.* Significantly, a *homeomorphism* or *isomorphism* between topological spaces, is a special case of homotopy equivalence, in which $g \circ f$ is equal (and not simply homotopic) to the *identity map* id_x and $f \circ g$ is equal to id_y. Therefore, if X and Y are homeomorphic, then they are homotopy-equivalent, but the opposite is not true.

the (topological) coalgebraic structures of a dissipative QFT. In this case, indeed, the DDF—or "active mirroring" (quantum entanglement) system-thermal bath—acts as a (thermo-)dynamic selection criterion of admissible sets, for the modal Boolean logic quantum computations of our "dissipative brains" [90].

On the other hand, it is easy in the light of the precedent discussion to guess that, just as a physical counterpart of the maximin principle in economy for a fair distribution of resources is in the "fair distribution of energy" among the components of a complex dissipative system balanced with its thermal bath, so Sen's *extended identity axiom* between different subjective basal spaces in SCS $(R_i, \tilde{R}_j)$, has in the DDF principle $(A, \tilde{A})$ its natural implementation (see Sect. D2 in Appendix D). This can be used as the basis of an unsupervised quantum ML algorithm inspired to the dissipative QFT underlying brain network dynamics (see Sect. B7 in Appendix B and [34]). Or—if we prefer to use the "first-person" jargon of the intentional language for expressing Smith's "extended sympathy" principle (see Sect. B1 in Appendix B and the connected diagram of cognitive sciences of Fig. 4)—, only by mirroring "myself" in "you" so to be each "the double" of the other, we can constitute a sympathetic "we".

Appendix D: A QFT Inspired Unsupervised Machine Learning Algorithm

D1. The Operator Algebra from Physics to Logic and Computer Science

In other contributions strictly related with the present one [34, 38, 65], we discussed at length the possibility of a fruitful modeling of the *intentional* behavior in human and artificial agents—strictly related to a "value based" deontic logic—, using the *dissipative* quantum field theory (QFT) approach to brain (thermo-)dynamics and to theoretical computer science (TCS). This modeling must be developed in the framework of a *topological approach to modal Boolean logic* [33], based on the momentous Marshall Stone's *Representation Theorem of Boolean Algebras* [55], from which a BAO directly derives [56, 57] that can be extended to *modal BAOs* [42, 43, 90], using the unifying framework of the Category Theory (CT) metalanguage (see [91] for a wider discussion).

What is fundamental for guessing, at least, the core of this passage from physics to logic, it is sufficient to recall Stone's powerful mathematical notion of *field of sets,* effectively a σ*-algebra,* that is a *probability space* in which a *metric* is defined, and that is typical of the physical system model theory, on which a BAO can be directly defined. All this can be resumed in the motto: *operator algebra from physics to logic*, disclosing an incredible panorama of development for *formal philosophy* (formal ontology, formal epistemology, formal ethics), on the one hand, and for *computational topology* in TCS, on the other hand. Included the actual growing discussion

about the development of *topological methods of statistical data management and of ML* [31]. See also [54] for a wider discussion of all the theoretical passages just sketched.

CT, indeed, is able to unify in the same axiomatic framework of the algebraic (topological) logic of *operator algebras*, both the *mathematical logic* of the natural sciences—of physics, before all—, and the *modal logic* of the philosophical disciplines [33, 58, 59, 91], the *deontic logic* of ethics included, with evident consequences for TCS and also for our problem of the *ethical accountability* of the AI algorithms and systems. The core of the CT logic (semantics) in its application to TCS is, indeed, the possibility of interpreting the *meaning function* [.], i.e., the function mapping a formula φ of the propositional calculus of a BAO [56, 57] over its extension $[\varphi]$ "making true" φ, not on a set-subset ordering like in standard set-theoretic semantics, but (primarily)[27] on a *complex algebra* $\mathbb{A}^+$, i.e., an algebra-subalgebras structure, so to satisfy the motto of CT logic, "meaning is a homomorphism" between algebraic structures [33].

Particularly, in Kripke's modal relational semantics in its algebraic (topological) interpretation, it is possible to justify:

1. A HOL semantics quantifying over *all* the truth valuation functions V for proposition p over world-states w: $\forall V(p, w)$.
2. Or a FOL semantics quantifying over all the possible states w of the world related with one state, i.e. over a *partition* (set disjoint union or *coproduct*) of the universe $\restriction(W)$. That is: $\forall w(V(p, w)|w \in \restriction (W)$ [89].

Indeed, in the CT metalanguage, Kripke modal relational semantics is defined over a coalgebra of trees of NWF-sets defined on Stone spaces [92]. More generally, by using the so-called "Vietoris transformation" as a selection criterion of admissible sets (set partitions or coproducts), it is possible justifying in CT, the *dual equivalence* between the category of coalgebras defined onto topological Stone spaces, and the category of modal Boolean algebras for the double *contravariant application* of the same "Vietoris functor" $\mathcal{V}/\mathcal{V}^*$, i.e., **SCoalg**$(\mathcal{V}) \simeq$ **MBAlg**$(\mathcal{V})$***** [62, 92]. Indeed, the core of a coalgebraic semantics of a Boolean algebra in CT logic is the Stone duality as a particular case categorical duality. Indeed, given the *dual equivalence* between a given category $\mathcal{C}$ and the opposed category $\mathcal{C}^{\text{op}}$, a statement α defined on $\mathcal{C}$ is true *only and only if* the opposed statement α^{op} defined on $\mathcal{C}^{\text{op}}$ is also true (see [93] for more details). In this sense, it is possible to develop a *relation ethics*, that is, a value-based deontic logic founded on a deontic interpretation of Kripke's modal relational semantics.

The connection with the mathematical formalism of QFT in the framework of the CT metalanguage is double, as we anticipated in Sect. 3.2. On the one hand, the topologies of the *Stone spaces* of the momentous "Stone Representation Theorem for Boolean Algebras" [55]—on which the extension of the operator algebra approach

[27] Indeed, by the application of the so-called "forgetful functor" it is always possible in CT mapping the category of monoids (one-object algebraic structures) **Mon** on the category of (pointed) sets **Set**, "forgetting" the underlying algebraic structure [93].

to Boolean logic is based in TCS [56, 57]—are the same of the topological spaces on which the *C*-(sub)algebras of Hilbert spaces* are defined, in the operator algebra formalism of QM and QFT [66, 94]. On the other hand, the other strong connection is based on the role of the *coproducts* (disjoint sums) and then of the *coalgebras* in quantum physics—effectively the coalgebras of the *Hopf bi-algebras* (algebra-coalgebra), systematically used in the calculations over lattices of quantum numbers, both in QM and QFT.[28] More precisely, coproducts play an essential role in QFT applied to *dissipative* systems modelled in *far-from-equilibrium* conditions because passing through different phases [95]. This interpretation inaugurated by the pioneering works of N. Bogoliubov [96, 97] and H. Umezawa [98–100], is the Fundamental Physics of *dissipative systems,* both in the relativistic quantum cosmology, and in the condensed matter physics, chemical and biological systems included [101]. The Bogoliubov transform, indeed, allows to map between different phases of the bosons and the fermions quantum fields, making QFT—differently from QM and from QFT in its Dirac's "second quantization" interpretation—able to calculate over *phase transitions.* One of the main differences between these two QFT modeling is that while the coproducts for calculating the *total energy* of a superposition quantum state are defined on a commutative algebraic footing because of the interchangeable character of the terms (superposed particles) of the quantum state, this commutativity does not hold in the dissipative case. In dissipative quantum systems this commutativity does not hold, because the two terms of the coproduct refer to the system and the thermal bath energy contributions that does not interchange each other and determining a *far-from-equilibrium balanced* quantum state. In this case, we are obliged to speak about *non-commutative q-deformed Hopf coalgebras*, where q is a thermal parameter, strictly related with the θ-angle of the Bogoliubov transform [101].

All this allows the possibility in CT logic to demonstrate the *dual equivalence* between the category of the non-commutative (*q-deformed*) *Hopf Coalgebras on Stone Spaces* **SHCoalg** of the dissipative QFT where the Bogoliubov transform in phsyics acts like the Vietoris transform in logic as a (dynamic) selection criterion of admissible sets for the coalgebraic semantics of the related BAO, and the category of the non-commutative ("skew") *modal Boolean algebras with operators* **MBAlg** for the contravariant application of the *Bogoliubov functor* $\mathcal{B}$, i.e., $\mathbf{SHCoalg}(\mathcal{B}) \simeq \mathbf{MBAlg}(\mathcal{B})^{\mathbf{*}}$ [54, 65].

This, on the one hand, offers an extension to quantum physics of the powerful and successful Jan Rutten's interpretation of the category of coalgebras as a *general theory of dynamic and computational systems,* both interpreted as *labelled state transition systems* (LTS) [102]. According to this theory, by applying the dual categorical equivalence algebras-coalgebras for the contravariant application of the same functor of CT logic, the dynamics of the physical system in which a Boolean logical calculus is implemented, as far as coalgebraically modelled, *directly gives* the semantics of the correspondent Boolean logical calculus in which the program is written.

[28] Effectively the Hopf coproducts are systematically used in QFT for calculating the total energy (sum) of n particles superposed in the same quantum state [104].

On the other hand, the non-commutative Hopf coalgebras (coproducts), through the powerful construction of the *doubling of the degrees of freedom* (DDF: see [101] and more synthetically [69] for a formal justification), satisfy a *dynamic* criterion of choice of admissible sets for justifying in CT logic a coalgebraic relational semantics of a modal Boolean algebra for Kripke models. Moreover, one of the most successful applications of dissipative QFT is for giving the Fundamental Physics of the mammalian brain dynamics, interpreted as a dissipative system (the "dissipative brain") [63, 64, 103]. This justifies a possible solution of the long-lasting problem in neurosciences of the "long-term memory traces", in terms of the coherent oscillatory behavior of large arrays of neurons also reciprocally very distant, in different areas of the mammalian brain and therefore that cannot be justified in terms of signals using synaptic paths. This *macroscopically* measurable behavior can have its only possible *microscopic* physical justification in terms of the *long-range correlations* (entanglement) of the quantum fields of the molecular components of the brain neuropile. These phase coherence neural domains, indeed, can coexist without interferences in the same *ground state* or "minimum energy condition" ("quantum vacuum condition (QV))" of a balanced state of the quantum fields, according to the powerful QFT construction of the *QV-foliation* $|0(\theta)_{\mathcal{N}}\rangle$ [65, 101]. This suggests the possibility of a *dynamic deep learning* strategy in artificial neural networks (ANN) and in AI systems both in the "supervised" [104], and in the "unsupervised" cases [65].

Moreover, from the cognitive neuroscience standpoint, all this demonstrates that thermal QFT is the Fundamental Physics of the mind conscious *intentionality* [50]. This foundation is consistent with Antonio Damasio's suggestion of interpreting the notion of *homeostasis* with the environment, based on complex non-linear self-regulation processes in biological and neural systems, as the physical basis of "individual" and "collective" *first-person intentionality* [39, 105], in its *third-person* or "observational" scientific modelling (biosemiotics) [38] (see Sect. B1 in Appendix B and Fig. 4).

D2. Applications of the DFF in QFT Unsupervised Learning

In the theoretical framework of thermal QFT for dissipative systems shortly discussed before, we sketch here a machine learning algorithm inspired to the DDF principle in dissipative QFT systems, in an optical ANN implementation, using the standard tools of the *correlation interferometry*, just as, for instance in the applications discussed in [106, 107]. In our case, indeed, the DDF principle can be applied in a recursive way, by using the *mutual information* as a measure of *phase distance,* like an optimization tool of error minimization of the input–output mismatch. In this case, indeed, the input of the net is not on the initial conditions of the net dynamics but on the boundary conditions (thermal bath) of the system. Just as it happens in the "deep learning" of natural brains, modelled as dissipative brains "locked" onto their environment variations (data streaming). In both cases, indeed, we are faced with the *macroscopic* phenomenon of *a dynamic phase locking*, having at the *microscopic* level in the DDF

principle of quantum entanglement between the degrees of freedom A of the system dynamics, and the degrees of freedom $\tilde{A}$ of its environment dynamics its proper explanation.

Indeed, inspired by the modeling of natural brains as many-body systems, the QFT dissipative formalism has been used to model ANNs [104, 108], also in the CT framework of a coalgebraic logic applied to TCS [65]. The QFT approach to brain studies was originally proposed by Ricciardi and Umezawa in 1967 [109], and extended in 1995 to include dissipative dynamics by Vitiello [90, 110]. The mathematical formalism of QFT (details in [101]) requires that for *open* (dissipative) systems, like the brain which is in a permanent "trade" or "dialog" with its environment, the degrees of freedom of the system (the brain), say A, need to be "doubled" by introducing the degrees of freedom $\tilde{A}$ describing the environment, according to the coalgebraic scheme: $A \rightarrow A \times \tilde{A}$. One is thus led to consider the deformed Hopf algebra, out of which *Bogoliubov transformations* involving the $A, \tilde{A}$ modes are derived. These transformations induce *phase transitions*, i.e., transitions through physically distinct spaces of the states describing different dynamical regimes in which the system can sit. The brain is thus continuously undergoing phase transitions (*criticality*) under the action of the inputs from the environment ($\tilde{A}$ modes). The brain activity is therefore the result of a continual balancing of fluxes of energy (in all its forms) exchanged with the environment. The balancing is controlled by the minimization of the free energy at each step of time evolution. Since fluxes "in" for the brain (A modes) are fluxes "out" for the environment ($\tilde{A}$ modes), and vice-versa, the $\tilde{A}$ modes are the *time-reversed images* of the A modes, they represent the *Double* of the system [90].

From the TCS standpoint this means that the system satisfies the notion of a particular type of automaton, or *Labelled State Transition Machine* (LTM). I.e., the so-called *infinite-state* LTM coalgebraically interpreted, and used for modelling *infinite streams of data* [102]. Effectively, also the QFT many-body systems are characterized by an infinite number of degrees of freedom. However, the doubling of the degrees of freedom (DDF) $\left\{A, \tilde{A}\right\}$ just introduced and characterizing a dissipative QFT system acts recursively as a *dynamic* selection criterion of admissible because balanced states (minimum of the free energy), and then as an unsupervised ML algorithm. Effectively, it acts as a mechanism of "phase locking" between the data flow (environment) and the system. Moreover, each system-environment entangled (doubled) state is *univocally* characterized by a *dynamically generated code*, or *dynamic labelling* as we see immediately. This means that this system is characterized by an "unlabeled" (memory recording) process, and these two properties—computing on data streaming, and performing an unsupervised learning—are the two main differences with the supervised learning algorithms illustrated in Appendix B.

In the model, indeed, an input triggers the *spontaneous breakdown of the symmetry* (SBS) of the system dynamical equations. As a result of SBS, massless modes, called Nambu-Goldstone (NG) modes, are dynamically generated [111, 112]. They are boson quanta of long-range correlations among the system elementary components and their *coherent condensation* in the system ground state (the least energy state or

"vacuum") describes the recording of the information carried by that input. *Coherence* denotes that the long-range correlations are not destructively interfering. Their macroscopic manifestations are the observable ordered patterns characterizing the system behavior. This is controlled by the order parameter, describing the degree and the specific nature of ordering. It is a *classical* field since, due to coherence, it does not depend on quantum fluctuations and the system is said to be a *macroscopic quantum system,* in the sense that its macroscopic dynamics and behavior is not derivable without recourse to the quantum dynamics.

The memory state turns out to be a *squeezed coherent state*: $|0(\theta)\rangle = \sum_j w_j(\theta)|w_j\rangle$ to which Glauber information entropy measure Q directly applies [113], with $|w_j\rangle$ denoting states of A and $\hat{A}$ pairs, θ is the time- and temperature-dependent Bogoliubov transformation parameter. $|0(\theta)\rangle$ is, therefore, a time-dependent state at finite temperature; it is an entangled state of the modes A and $\hat{A}$, which provides the mathematical description of the *unavoidable* interdependence between the brain and its environment. Coherence and entanglement imply that quantities relative to the A modes depend on corresponding ones of the $\tilde{A}$ modes. From the CT logic standpoint, this means that a "truth evaluation function" is built-in the $\left\{A, \hat{A}\right\}$ system.

More analytically, The Bose–Einstein distribution function of the A_k and $\tilde{A}_k$ modes is determined by the minimization of the free energy: $N_k = 1/\left(e^{\beta\omega} - 1\right)$, with N_k the number of condensed A modes, $\beta = 1/k_B T$, k_B the Boltzmann constant, $\omega = \omega_k$ the energy (and similarly for $\tilde{N}_k$). For simplicity, we write $\theta = \theta(t, \beta(t))$, omitting dependence of θ on time t and temperature T. The collection $\left\{N_k, \tilde{N}_k; N_k - \tilde{N}_k = 0, \text{for any k}\right\}$ acts as a *code* (a *dynamically generated label*) associated with the information *printed* by the condensation of the $\left\{A_k, \tilde{A}_k\right\}$ in $|0(\theta)\rangle$.

In the presence of fermion fields, SBS also leads to the formation of NG boson condensation modes, with their Bose–Einstein distribution functions. The fermion modes are also doubled and $|0(\theta)$ is the tensor product of the (Bogoliubov transformed) fermion states and of the NG boson states. The fermion number in the fermion state is given by the Fermi–Dirac distribution function: $NF_k = 1/\left(e^{\beta\omega} + 1\right)$, for any k, and similarly for the tilde-fermion mode. Here, as usual, $\omega = \omega_k = \varepsilon_k - \mu$, with $\omega = \omega_k = \varepsilon_k$ the energy and μ the chemical potential.

We stress that in the QFT dissipative formalism the implicit discreteness in the *on–off* (or 0/1) algebra, although it may be present at a microscopic level—e.g., in the fermion and dipole quantum numbers, etc.—it is *dynamically* converted (through the *dynamical rearrangement of symmetry*) into a *continuous* interval [0, 1] of probability values. In fact, from the Bose–Einstein and Fermi–Dirac distributions, one can derive *the sigmoid activation function* σ. For example, in the fermion case, assuming that $N_{Fk} = 1 \text{at } T = 0$ and energy $\varepsilon_k < \mu$, one finds that the change ΔN_{Fk} due to thermal effects is given by: $\Delta N_{Fk} = 1 - 1/\left(e^{\beta\omega} + 1\right) = 1/\left(e^{\beta\omega} + 1\right) = \sigma$, the sigmoid function, indeed (formal details in [104, 108]). In the boson case, considering that $N_k = \sinh^2\theta \text{and } e^{-\beta\omega} = \tanh^2\theta$, it is also not difficult to describe the system

response in terms of the sigmoid function σ. However, we must emphasize that in the present case the value ("slope") of σ is not "put by hand", but like the "labels" N_{k}, $N_{F\mathrm{k}}$ depends on the system *dynamics*, as the strict relationship between these two magnitudes in the defining formulas above demonstrates.

References

1. Floridi, L., Taddeo, M.: What is data ethics. Phil. Trans. Roy. Soc. A: Math., Physic., Engineer. Sc., **374**(2083), 20163060 (2016)
2. Basti, G.: Ethical responsibility vs. ethical responsiveness in conscious and unconscious communication agents. Proceedings **47**(68), 1–7 (2020)
3. Libet, B., et al.: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act. Brain **106**(3), 623–642 (1983)
4. Haynes, J., Roth, G., Stadler, M., Heinze, H.: Reading intentions in the human brain. Curr. Biol. **17**(4), 323–328 (2007)
5. Basti, G.: Filosofia dell'uomo. ESD, Bologna (2010)
6. Levy, N.: Consciousness and Moral Responsibility. Kindle edition, Oxford UP, Oxford UK (2014)
7. Gazzaniga, M.S.: Who Is in Charge? Free Will and the Science of the Brain. Harper Collins Publ, New York (2011)
8. Levy, N.: Neuroethics: Challenges for the 21st Century, Cambridge UP, Cambridge UK (2007)
9. Müller, V.C.: Ethics of Artificial Intelligence and Robotics, 1 June 2021, https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/. Last accessed 2022/11/10
10. D'Amodio, A.: Toward a human-centered economy and politics: the theory of justice as fairness from Rawls to Sen. Philosophies **5**(4, 44), 1–50 (2020)
11. Whittaker, M., et al.: AI Now Report 2018, AI Now Institute, New York University, 2018, https://ainowinstitute.org/AI_Now_2018_Report.html. Last accessed 2022/01/27
12. Christman, J.: Autonomy in moral and political philosophy. In: Zalta, E. (ed.) Stanford Encyclopedia of Philosophy, (Fall 2020 Edition), 1 October 2020, https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/. Last accessed 2022/12/28
13. Dignum, V.: Ethics in artificial intelligence: introduction to the special issue. Ethics Inf. Techn. **20**(1), 1–3 (2018)
14. Sen, A.: Collective Choice and Social Welfare, Expanded Penguin Ltd., Kindle Edition, London (2017)
15. Wiener, N.: Cybernetics. Or Communication and Control in the Animals and the Machines, 2nd edn. MIT Press, Cambridge, MA (1962)
16. Turing, A.M.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
17. McCarthy, J., Minsky, M., Rochester, N., Shannon, C.: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1 August 1955, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf. Last accessed 2023/01/01
18. Damasio, A.: Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Publishing, New York (1994)
19. Benzmüller, C., Parenta, X., van der Torre, L.: Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. Artif. Intell. **287**(103348), 1–50 (2020)
20. Lo Piano, S.: Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit. Soc. Sci. Commun. **7**(9), 1–7 (2020)
21. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (2018)

22. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford UP, Oxford, UK (2017)
23. Gajane, P., Pechenizkiy, M.: On Formalizing Fairness in Prediction with Machine Learning, 28 May 2018, https://arxiv.org/abs/1710.03184v3. Last accessed 2023/01/05
24. Card, D., Smith, N.A.: On consequentialism and fairness. Front. Artif. Intell. **3**(34), 1–11 (2020)
25. Benzmüller, C., Sultana, N., Paulson, L.C., Theiss, F.: The higher-order prover LEO-II. J. Aut. Reas. **55**(4), 389–404 (2015)
26. Benzmüller, C., Andrews, P.: Church's type theory. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy, May 2019 edn. https://plato.stanford.edu/entries/type-theory-church/. Last accessed 2023/10/01
27. Birhane, A., Cummins, F.: Algorithmic injustice: toward a relational ethics, https://arxiv.org/pdf/1912.07376v1.pdf. Last accessed 2022/12/10 (2019)
28. Přibáň, J. (ed.): Liquid Society and its Law. Routledge, New York (2020)
29. Endriss, U.: Logic and social choice theory. In: Gupta, A., Van Benthem, J. (eds.) Logic and Philosophy Today, pp. 333–377. College Pubblications, London (2011)
30. Sen, A.K.: The possibility of social choice. Nobel Lecture. Am. Econ. Rev. **89**(1), 178–215 (1999)
31. Hensel, F., Moor, M., Rieck, B.: A survey of topological machine learning methods. Front. Art. Intell. **4**(681108), 1–12 (2021)
32. Aczel, P.: Non-Wellfounded Sets, CLSI Lecture Notes, vol. 14, Stanford UP, Stanford CA (1988)
33. Venema, Y.: Algebras and co-algebras. In: Blackburn, P., van Benthem, F.J.F., Wolter, F. (eds.) Handbook of Modal Logic, pp. 331–426. Elsevier, Amsterdam (2007)
34. Basti, G., Capolupo, A., Vitiello, G.: The computational challenge of Amartya Sen's social choice theory in formal philosophy. In: Giovagnoli, R., Lowe, R. (eds.) The Logic of Social Practices. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 52. Springer, Berlin-New York, pp. 87–119 (2020)
35. Sen, A.K.: The Idea of Justice. Penguin Books Ltd., London, UK (2010)
36. Sen, A.K.: The impossibility of a Paretian liberal. J. Pol. Econ. **78**(1), 152–157 (1970)
37. Arrow, K.J.: Social Choice and Individual Values, 2nd edn. Yale UP, New Haven & London (1963)
38. Basti, G., Capolupo, A., Vitiello, G.: The doubling of the degrees of freedom in quantum dissipative systems, and the semantic information notion and measure in biosemiotics. Proceedings **47**(69), 1–7 (2020)
39. Damasio, A.: The Strange Order of Things. Life, Feeling and the Making of Cultures. Pantheon Books, New York (2018)
40. Baird, A., Schuller, B.: Considerations for a more ethical approach to data in AI: on data representation and infrastructure. Front. Big Data **3**(25), 1–11 (2020)
41. Cresswell, M.J., Huges, G.E.: A New Introduction to Modal Logic. Routledge, London (1996)
42. Galvan, S.: Logiche intensionali. Sistemi proposizionali di logica modale, deontica, epistemica. Franco Angeli, Milano (1991)
43. Hoftstadter, D.: Gödel, Escher, Bach: An Eternal Golden Braid, 1st edn. Vintage Books, New York (1979)
44. Putnam, H.: Minds and Machines. Dimensions of Mind, Collier, New York (1960)
45. Gardner, H.: The Mind's New Science: A History of the Cognitive Revolution. Basic Books, New York (1984)
46. Feigl, H.: The "mental" and the "physical". In: Feigl, H., Scriven, M., Maxwell, G. (eds.) Minnesota Studies in the Philosophy of Mind. Vol. II: "Concepts, Theories and the Mind-Body Problem", Minnesota UP, Minneapolis, pp. 370–497 (1958)
47. Sellars, W.: Intentionality and the mental. In: Feigl, H., Scriven, M., Maxwell, G. (eds.) Minnesota Studies in the Philosophy of Mind. Vol. II: "Concepts, Theories and the Mind-Body Problem", Minnesota UP, Minneapolis, pp. 507–524 (1958)

48. Quine, W.V.O: Quiddities. An Intermittently Philosophical Dictionary, Harvard UP, Cambridge MA (1987)
49. Shannon, C.: The mathematical theory of communication. Bell Syst. Techn. J. **27**(3), 379–423 (1948)
50. Freeman, W.J.: Intentionality. Scholarpedia **2**(2), 1337 (2007)
51. Searle, J.R.: Minds, brains, and programs. Behav. Brain Sci. **3**, 128–135 (1980)
52. Searle, J.R.: Intentionality. An Essay in the Philosophy of Mind. Cambridge UP, New York (1983)
53. Dreyfus, H.: Husserl's perceptual noema. In: Dreyfus, H. (ed.) Husserl, Intentionality and Cognitive Science, pp. 97–124. MIT Press, Cambridge, MA (1982)
54. Basti, G.: The philosophy of nature of the natural realism. The operator Algebra from physics to logic. Philosophies **121**(7), 1–84 (2022)
55. Stone, M.H.: The theory of representation for Boolean algebras. Trans. Amer. Math. Soc. **40**, 37–111 (1936)
56. Jónsson, B., Tarski, A.: Boolean algebras with operators. Part I. Am. J. Math. **73**, 891–939 (1952)
57. Jónsson, B., Tarski, A.: Boolean algebras with operators. Part II. Am. J. Math. **74**, 127–152 (1952)
58. Goldblatt, R.I.: Metamathematics of modal logic I. Rep. Math. Log. **6**, 41–48 (1976)
59. Goldblatt, R.I.: Metamathematics of modal logic II. Rep. Math. Log. **7**, 21–52 (1976)
60. Kripke, S.A.: Semantical analysis of modal logic I. Normal modal propositional logic calculi. Zeitschr. Math. Log. Grundl. Math. **9**, 67–96 (1963)
61. Kripke, S.A.: Semantical analysis of modal logic II. Non-normal modal propositional calculi. In: Addison, J.W., Henkin, L., Tarski, A. (eds.) The Theory of Models, pp. 206–220. North Holland, Amsterdam (1965)
62. Abramsky, S.: A Cook's tour of the finitary non-well-founded sets (original lecture: 1988). In: Artemov, S., Barringer, H., d'Avila, A., Lamb, L.C., Woods, J. (eds.) Essays in Honor of Dov Gabbay, vol. I, pp. 1–18. Imperial College Pubblications, London (2005)
63. Freeman, W.J., Vitiello, G.: Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics. Phys. Life Rev. **3**(2), 93–118 (2006)
64. Freeman, W.J., Vitiello, G.: Dissipation and spontaneous symmetry breaking in brain dynamics. J. of Phys. A: Math. Theor. **41**(30), 304042 (2008)
65. Basti, G., Capolupo, A., Vitiello, G.: Quantum field theory and coalgebraic logic in theoretical computer science, Prog. in Bioph. & Mol. Biol. Special Issue: Quantum information models in biology: from molecular biology to cognition. **130**(A) 39–52 (2017)
66. Landsman, K.N.P.: Foundations of Quantum Theory. From Classical Concepts to Operator Algebra. Springer, Berlin-New York (2017)
67. Kozma, R., Puljic, M., Freeman, W. J.: Thermodynamic model of criticality in the cortex based on EEG/ECoG data. In: Plenz D., Niebur, E. (eds.) Criticality in Neural Systems. Wiley-VCH Verlag GmbH & Co. KGaA Hoboken NJ, pp. 153–176 (2014)
68. Fodor, J.A.: Modularity of Mind: An Essay on Faculty Psychology. MIT Press, Cambridge MA (1983)
69. Pylyshyn, Z.W.: Computation and Cognition: Toward a Foundation for Cognitive Science. MIT Press, Cambridge MA (1986)
70. Minsky, M.: The Society of Minds. Simon & Schuster, New York (1986)
71. McCulloch, W.S., Pitts, W.H.: A logical calculus of ideas immanent in nervous activity. Bull. Math. Biophys. **5**, 115–133 (1943)
72. Hebb, D.O.: The Organization of Behavior. Wiley & Sons, New York (1949)
73. Löwel, S., Singer, W.: Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. Science **255**, 209–212 (1992)
74. Farley, B.G., Clark, W.A.: Simulation of self-organizing systems by digital computer. IRE Trans. Inform. Th. **4**, 76–84 (1954)
75. Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Cornell UP, Buffalo NY (1961)

76. Minsky, M., Papert, S.: Perceptrons. An Introduction to Computational Geometry, 2nd edn. MIT Press, Cambridge, MA (1987)
77. Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., Debashis, D.: Fundamental concepts of convolutional neural network. In: Balas, V.E., Kumar, R., Srivastava, R. (eds.) Recent Trends and Advances in Artificial Intelligence and Internet of Things. Intelligent Systems Reference Library, vol. 172. Springer, Berlin, New York, pp. 519–567 (2020)
78. Freeman, W.J.: How Brains Make Up Their Minds. Columbia UP, New York (2001)
79. Minsky, M.: Frame theory. In: Johnson-Laird, P.N., Wason, P.A. (eds.) Thinking: Reasings in Cognitive Science, pp. 355–376. Cambridge UP, Cambridge, MA (1977)
80. Rumelhart, D., PDP Group: Parallel Distributed Processing. Voll. 1–2. MIT Press, Cambridge, MA (1986)
81. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
82. Werbos, P.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard UP, Boston, MA (1974)
83. Basti, G., Perrone, A. L.: Chaotic neural nets, computability, undecidability. An outlook of computational dynamics. Int. J. Int. Syst. **10**(1), 41–69 (1995)
84. Basti, G., Vitiello, G.: A QFT approach to data streaming in natural and artificial neural networks. Proceedings **81**(1), 106 (2022)
85. Vitiello, G.: Classical chaotic trajectories in quantum field theory. Int. J. Mod. Phys., B **18**, 785–792 (2004)
86. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems—vol. 1. ACM Publ., New York, pp. 1097–1105 (2012)
87. Stanford Vision Lab., Stanford University and Princeton University: Imagenet Website and Dataset (Update), 11 March 2021. https://www.image-net.org/update-mar-11-2021.php. Last accessed 2023/01/08
88. Kehrenberg, T., Chen, Z., Quadrianto, N.: Tuning fairness in balancing target labels. Front. Artif. Intell. **3**(33), 1–12 (2020)
89. Goranko, V., Otto, M.: Model theory of modal logic. In: Blackburn, P., van Benthem, F.J.F., Wolter F. (eds.) Handbook of Modal Logic. Elsevier, Amsterdam, pp. 252–331 (2007)
90. Vitiello, G.: My Double Unveiled. John Benjamins Publ. Co., Amsterdam (2001)
91. Goldblatt, R.I.: Topoi: The Categorial Analysis of Logic, Revised Elsevier, Amsterdam (1984)
92. Kupke, C., Kurz, Venema, Y.: Stone coalgebras. Theoret. Comp. Sci. **327**(1), 109–134 (2004)
93. Abramsky, S., Tzevelekos, N.: Introduction to categories and categorical logic. In: Coecke, B. (ed.) New Structures for Physics. Lecture Notes in Physics, vol. 813, Springer, Berlin-New York, pp. 3–94 (2011)
94. Landsman, K.N.P.: Lecture notes on operator algebras, 14 December 2011, http://www.math.ru.nl/~landsman/OA2011.html. Last accessed 2023/01/10
95. Becchi, C.M.: Second quantization. Scholarpedia **5**(6), 7902 (2010)
96. Bogoliubov, N.N.: On a new method in the theory of superconductivity. Nuovo Cimento **7**(6), 794–805 (1958)
97. Bogoliubov, N.N., Tolmachev, V.V., Shirkov, D.V.: A New Method in the Theory of Superconductivity. Chapman & Hall, New York, London (1959)
98. Takahashi, Y., Umezawa, H.: Thermo-field dynamics. Coll. Phen. **2**, 55–80 (1975)
99. Umezawa, H.: Advanced Field Theory: Micro, Macro and Thermal Concepts. American Institute of Physics, New York (1993)
100. Umezawa, H.: Development in concepts in quantum field theory in half century. Math. Japonica **41**(1), 109–124 (1995)
101. Blasone, M., Jizba, P., Vitiello, G.: Quantum field theory and its macroscopic manifestations. Boson Condensations, Ordered Patterns and Topological Defects. Imperial College Press, London (2011)
102. Rutten, J.J.M.: Universal coalgebra: a theory of systems. Theor. Comp. Sci. **249**(1), 3–80 (2000)

103. Vitiello, G.: The dissipative brain. In: Globus, G.G., Pribram, K.H., Vitiello, G. (eds.) Brain and Being—At the Boundary Between Science, Philosophy, Language, and Arts, pp. 317–330. John Benjamins Pub. Co., Amsterdam (2004)
104. Pessa, E., Vitiello, G.: Quantum dissipation and neural net dynamics. Biochem. Bioenerg. **48**, 339–342 (1999)
105. Damasio, A.: Self Comes to Mind: Constructing the Conscious Brain. Heinemann, London (2010)
106. Basti, G. Bentini, G.G., Chiarini, M., Parini, A., Artoni A., et al.: Sensor for security and safety applications based on a fully integrated monolithic electro-optical programmable microdiffractive device. In: Proceedings of SPIE 11159, Electro-Optical and Infrared Systems: Technology and Applications XVI, SPIE Pub., Strasbourg (2019)
107. Parini, A., Chiarini, M., Basti, G., Bentini, G.G.: Lithium niobate-based programmable microdiffraction device for wavelength-selective switching applications. In: Proceedings of SPIE 11163, Emerging Imaging and Sensing Technologies for Security and Defence IV, SPIE Pub. Strasbourg (2019)
108. Vitiello, G.: Neural networks and many/body systems. In: Minati G. (ed.) Multiplicity and Interdisciplinarity. Essays in Honor of Eliano Pessa. Springer, Berlin-New York (2021)
109. Ricciardi, L.M., Umezawa, H.: Brain physics and many-body problem. Kibernetik **4**(1), 44–48 (1967)
110. Vitiello, G.: Dissipation and memory capacity in the quantum brain model. Int. J. Mod. Phys. B **9**, 973–989 (1995)
111. Nambu, Y.: Quasiparticles and gauge invariance in the theory of superconductivity. Phys. Rev. **117**, 648–663 (1960)
112. Goldstone, J.: Field theories with superconductor solutions. Nuovo Cimento **19**, 154–164 (1961)
113. Keitel, C.H., Wodkiewicz, K.: Measuring information via Glauber's Q-representation. In: Proceedings of the Second International Workshop on Squeezed States and Uncertainty Relations (1993)
114. Suppes, P.: Some formal models of grading principles. Synthese **16**(3–4), 284–306 (1966)
115. McCulloch, W.S., Pitts, W.H.: A logical calculus of ideas immanent in nervous activity. Bull. Math. Bioph. **5**(4), 115–133 (1943)
116. Rawls, J.: A Theory of Justice, Revised Belknap Press, Cambridge, UK (1999)
117. Weisskopf, V.F.: Knowledge and Wonder: The Natural World as Man Knows It, 2nd edn. MIT Press, Cambridge, MA (1979)
118. Quine, W.V.O.: Mathematical Logic, Revised edn. Harvard UP, Cambridge, MA, London (1981)
119. Zalta, E.: Intensional Logic and the Metaphysics of Intentionality. MIT Press, Cambridge, MA (1988)
120. Rawls, J.: Justice as fairness. Phil. Rev. **67**(2), 164–194 (1958)
121. Lewis, C.I., Langford, C.H.: Symbolic Logic, 2nd edn. Dover Publications, New Yok, 1959 edn. Century Company, New York (1932)