proceedings

Proceedings

Ethical Responsibility vs. Ethical Responsiveness in Conscious and Unconscious Communication Agents †

Gianfranco Basti

Faculty of Philosophy, Pontifical Lateran University, 00120 Vatican City, Italy; basti@pul.va;
Tel.: +39-339-576-0314

† Workshop Hacking Societies, Habits and Rituals, Berkeley, California, USA, 2–6 June 2019.

Published: 14 May 2020

Abstract: In this contribution, I start from Levy's precious suggestion about the neuroethics of distinguishing between "the slow-conscious *responsibility*" of us as persons, versus "the fast-unconscious *responsiveness*" of sub-personal brain mechanisms studied in cognitive neurosciences. However, they are both *accountable* for how they respond to the environmental (physical, social, and ethical) constraints. I propose to extend Levy's suggestion to the fundamental distinction between "moral responsibility of conscious communication agents" versus the "ethical responsiveness of unconscious communication agents", like our brains but also like the AI decisional supports. Both, indeed, can be included in the category of the "sub-personal modules" of our moral agency as persons. I show the relevance of this distinction, also from the logical and computational standpoints, both in neurosciences and computer sciences for the actual debate about an ethically accountable AI. Machine learning algorithms, indeed, when applied to automated supports for decision making processes in several social, political, and economic spheres are not at all "value-free" or "amoral". They must satisfy an ethical responsiveness to avoid what has been defined as the unintended, but real, "algorithmic injustice".

Keywords: neuroethics; digital ethics; quantum field theory

1. Introduction

In our information age, humans and machines are depending on each other ever more strongly, like many "conscious" and "unconscious" communication agents [1]. This opens the way to a systematic comparison between the moral "responsiveness" of natural (brains) and artificial (AI systems) *unconscious* communication agents, on the one hand, and the moral "responsibility" of their *conscious* human owners/users/designers, on the other hand, with respect to the common social/linguistic environment with its unavoidable ethical constraints [1,2].

Now, the main problem at issue when we discuss the actual challenges in moral philosophy related to *Neuroethics* (NE) and to *Artificial Intelligence* (AI) applied to autonomous systems is the following. "Who is the actor of a moral act?", or, by using the title of a famous book by Michael S. Gazzaniga on neuroethics, "who is in charge?" [3]. In this contribution, we suggest a historical and theoretical reconstruction of the main lines of this debate, of which it is difficult to ignore the relevance.

2. The Neuroethics Issue and the Challenge of Developing Ethically "Good" Algorithms in AI Systems

One of the fundamental contributions of cognitive neurosciences to overcoming the early *functionalist* or symbolic approach in cognitive science [4,5] is related to the development of *Neuroethics*, originally inspired by Antonio Damasio's criticism against the rationalist interpretation of the self-conscious mind by Descartes and Kant. This interpretation, indeed, is strictly related to the

representationalism of the functionalist approach to cognitive science by Fodor and Pylyshyn, reinterpreting Descartes' mind–body dualism in terms of the software–hardware dualism [5,6], with the consequent *individualism* depending on Carnap's "methodological solipsism" characterizing such an interpretation of the self, identified with the subjective self-consciousness, that Fodor explicitly extended to cognitive sciences [7].

Damasio's criticism against Descartes' dualism of mind and body and then of intelligence and emotion originated from the experimental evidence he collected during the 1990s, according to which, due to a defect in the *emotional area* of their brains, intelligent patients, without attention, language, and abstract reason defects, are systematically *unable to make decisions* that are profitable for their own persons, acceptable for their community, and then *morally correct* [8]. This evidence is a straightforward experimental falsification of Descartes' and the overall Kantian *rationalistic* foundation of ethics that criticized the traditional *intentional* foundation because it is based, according to Kant, on a *hypothetical* (conditional) argument "if you will *X*, you must do *Y*", for a *formalistic* foundation of morality based on an *apodictic* (unconditional) argument, and then on the abstract *universality* of the moral duty. That is, the so-called, "categorical imperative"[1], which is based on the "perfect duty", namely, the *unconditional* (tautological) "duty-for-duty", characterizing human *free will* or "practical reason". Human will as such has no determined content but it is determined by the universal *unconditional* because of tautological moral laws *that reason gives to itself*[2], as opposed to the *conditional* "duty-for-a-goal" characterizing sensible desires and impulses, which are determined by something else, the desired object, so characterizing the animal choices.

On the contrary, in the personalist approach to the intentional foundation of ethics, human free-will is interpreted as the intentional faculty of desiring something made *moral* or *ethically good*, by the "goodness of the desired goal", because fulfilling a *person's subjective and inter-subjective capabilities* and then granting *her/their happiness* depends ultimately on the person flourishing, where "person" is not the simple "individual", but the "individual-in-relationship".

It is evident that Damasio's empirical discoveries are compatible with such an intentional approach to the foundation of ethics, falsifying from the neuroscience standpoint Descartes' and Kant's ethical rationalism. Effectively, Damasio explicitly joined the intentional approach to epistemology and ethics in his more recent books devoted to the development of his early approach, also constituting the background of the actual neuroethics research program [10,11]. In fact, he vindicates that the intentional "aboutness", i.e., the *purposeful subject–object* relationship, against the subjectivist solipsism of Descartes' identification of the self with the self-consciousness, concerns primarily the biology before that the psychology. In fact, all living organisms satisfy a *homeostatic* two-way relationship with their environments. In brains, there exists *a two-way relationship* between brain and body and, through the body surface, with the outer physical and human environment, in order to satisfy the biological and social goals of the human individual(s) ([10], pp. 91–108), and on which the same construction of cultures ultimately depends [11]. In other terms, "intentionality" is not only *psycho*-logical, but primarily "*bio*-logical". On this homeostatic basis, Damasio refined the triangulation of cognitive neuroscience into the distinction among: (1) the *first-person awareness* he defined also as the "self-as-witness" or the subjective "presence-to-herself" of each human person; (2) the *self-consciousness*, as the always necessarily partial self-objectification to ourselves of our witnessing self; (3) the *moral behavior attribution* also to someone else of our moral agency capability, i.e., the third-person attribution of a moral *conscience* (aware evaluation of the goal–means relationship) to other humans, both individuals and groups. Damasio, therefore, individuated in the passage from (1) to (3) the *proper neuroethics research*, having in the *person*, i.e., "the aware individual-in-relationship with its natural/social environment" the *moral agent*, because of the irrelevance of (2), erroneously identified by Descartes as "the self" and then as the first-person "I", subject of the human moral agency.

---

[1] It has its more famous formulation in Kant's *Grundlegung zur Metaphysik der Sitten* (1785): "Act only according to that maxim whereby you can at the same time will that it should become a universal law" ([9], p. 30).

[2] "The will of every rational being is a universally legislating will".

From this standpoint, it is easy to understand which is the symmetrical error of *reductionism* as to Descartes' mistake of *dualism.*, that is, the attribution of the role of behavior controller to the brain isolated by its bodily and environmental surroundings. This mistake, as it is well-known, is based on the neurophysiological evidence that any conscious volition is causally determined by the readiness potentials of the neural modules involved in the production of some sensory-motor action. They precede by milliseconds (Bernard Libet) [12], or by tenths of seconds (John-Dylan Haynes) [13], the conscious expression of making a voluntary act, so that, using the title of the above quoted paper by Haynes and his collaborators, "reading hidden intentions in the human brain" is possible. From this and other similar evidence, somebody can arrive at the wrong conclusion that is symmetrically opposed to Descartes' mistake, reported by Michael S. Gazzaniga in his book on neuroethics that "conscious volition, the idea that you are willing an action to happen, is an illusion"([3], p. 129). Whereas, on the contrary, it is evident that the target of neuroscience criticism is the modern identification, from Descartes to Kant, of the self, as moral agent, with the *self-consciousness,* and the connected absurdity of the *causal role of consciousness* in human agency**.** The symmetrical mistake of identifying in the brain the "true" agent of our choices derives from the false supposition of considering the brain as an isolated agent.

> The point is that we now understand that we have to look at the whole picture, a brain in the midst of and interacting with other brains, not just one brain in isolation (…). Human nature remains constant, but out in the social world behaviour can change. Brakes can be put on unconscious intentions ([3], p. 216).

In other terms, what Gazzaniga emphasizes is that, in the light of neurosciences, responsibility and accountability, like the same term "responsible" (responding to someone else) highlights, has primarily a social and hence biological nature that is not opposed at all to the *personal* nature of a moral act. It is opposed only to an individualistic *monism* and/or *dualism*, which are the two opposed ideologies that contaminated the Western thought for millennia and against which neurosciences can give an essential decontamination contribution.

To sum up, we as persons are our brains and our bodies in relationship with our (physical–social) environment. Our conscious behavior and our awareness is only the tip of the iceberg, rather than the whole iceberg of our *free* moral agency. In this context, we can redefine *freedom* as the faculty of each human person as a whole, including their environmental relationships, of "self-determining their own behavior in view of effectively reaching some goal", where "goal" is synonymous with "an aware end", and the morality of a free act depends on the ethical and personalistic but non-individualistic character of the pursued goal (see [14], chp. 5 for a synthesis). To conclude with the words of another representative of the neuroethics approach, Neil Levy,

> 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 ([15], Kindle pos. 395–396).

Of course, among the "sub-personal mechanisms" for Levy, we have to consider primarily the neural module computations studied by cognitive neurosciences. It is important to quote Levy's position on the neuroethics debate because in a further book he introduces a precious distinction that is fundamental for our aims, that is, the distinction between "the slow-conscious *responsibility*" of us as persons versus "the fast-unconscious *responsiveness*" of unconscious skilled agents, both anyway *accountable* for responding to the environmental (physical, social, and ethical) constraints.

> It is characteristic of conscious processes that they are much slower than nonconscious; the rapid responsiveness of highly skilled agents (…) 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 ([16], p. 114).

To give full intelligibility to this very important distinction, I suggest to substitute the last "responsible" occurrence, with "accountable", to emphasize the fundamental distinction between "moral responsibility of conscious communication agents" and the "ethical responsiveness of unconscious communication agents", like our brains but also like the AI decisional supports, which both can be included in the category of the "sub-personal modules" of our moral agency.

The extension of a biological notion such as "responsiveness" that per se denotes the fast response of a healthy cell to its chemical environment brings us back to Damasio's evocative suggestion of extending the notion of intentionality as homeostasis to the social environment, or to the "society of brains", if we prefer to continue in a metaphorical usage of language. Indeed, in biology and in neuroscience, the notion of "homeostasis" says much more than the thermodynamic notions of "equilibrium", which in biology is synonymous with "death" ([11], p. 49) and with "energy balance" because it expresses complex strongly non-linear phenomena of "self-regulation", ranging from the molecular, to cellular, to tissue levels, and so on, eventually involving the whole organism and its environment (see the I Part of [11], pp. 11–69 for a review of this notion). Therefore, Walter Freeman's experimental demonstration of the chaotic character of brain dynamics in performing intentional tasks [17,18] is one of the best candidates to make Damasio's suggestion of a biological foundation of intentionality scientifically operational. Overall, if we follow Freeman and his collaborators in identifying the fundamental physics of the chaotic brain dynamics in the quantum field theory (QFT) of dissipative systems, in far-from-equilibrium conditions [19–21]. Indeed, what we observe at the *macroscopic* level as a chaotic trajectory in the brain phase space corresponds mathematically at the *microscopical* level to a trajectory between different phases of the quantum fields of the neuropil in human neo-cortices, "entangled" with their inner (body) and outer environments [19,20]. In this way, Damasio's suggestion of the biological foundation of intentional "aboutness" as homeostasis can find its proper operational modeling ([19], pp. 100–109).

Without the possibility of deepening here as I did elsewhere the informational [22] and hence computational [21] and logical [1,2,23] consequences of such an approach for modeling the natural and the artificial quantum neural computations, it is sufficient to recall here two significant points as to our precedent discussion.

First of all, the quantum entanglement characterizing the formation–dissolution of phase coherence domains of the electromagnetic fields in the brain, at different levels of matter organization, grants the fastest responsiveness to stimuli of neural modules that is available in nature. This implies that, on the one hand, as Freeman himself emphasized, the few milliseconds of the time that lapse between the stimulus arrival and its recognition in the neural dynamics cannot be granted in principle by the "gradient descent" operations of AI machine learning ([19], p. 107). On the other hand—and very significantly—the *hyperbolic tangent function,* which is the core of any non-linear weight activation function of artificial neural network (ANN) dynamics able to reckon with the higher order correlations in the data set [24], is also the essential component of the so-called "Bogoliubov transform", which generally controls the photon (boson) condensate phase transitions in the QFT of condensed matter systems – the biological and the neural systems included [21,25]. This suggests a QFT approach to the continuous redefinition of the neural net weights, both in natural and artificial neural systems, using this type of quantum "deep learning", to deal with the always changing higher order correlations in the data streaming from the environment [21].

3. Conclusions: The Necessary Path toward an Ethically Accountable AI

The second conclusive point to emphasize is that this intentional approach to cognitive neurosciences based on the quantum *entanglement* brain–environment acquires an immediate logical and computational relevance if we consider that the same *coalgebraic* modeling [26] holds, both for "dissipative" QFT systems in the foundations of condensed matter physics [25]—biological [27–30] as well neural systems included [19,20]—and for Saul Kripke's *modal relational semantics* in theoretical computer science [31–34][3]. The modal calculus is, indeed, the proper logic of formal ontology, but

[3] (Co-)Algebraic coproducts, indeed, are effectively sums. Therefore, in modelling physical dissipative systems in QFT it is necessary to use non-commutative coproducts and the consequent "algebra doubling":

also of formal ethics, and of their algorithmic applications [32]. This is, of course, relevant for the debate that is recently growing about an ethically accountable AI in order to avoid what has been defined—in a provocative contribution presented at the "Black in AI workshop" during the *NIPS 2019-Neural Information Processing System Conference* [35]—as the unintended but real "algorithmic injustices". Avoiding these problems requires, indeed, according to the Authors of the contribution, "developing and deploying ethical algorithmic systems", in the context of a "relational approach to ethics". Effectively, with the standard statistical approach to machine learning, "the systems 'pick-up' social and historical stereotypes (…), individuals and groups often at the margins of society that fail to fit stereotypical boxes, suffer the undesirable consequences" ([35], p. 1).

This means that AI machine learning algorithms, when applied to automated supports for decision making processes in several social, political, and economic spheres, are not at all "value-free" or "amoral", but they must satisfy an ethical responsiveness. Now, what the authors intend by the "relational ethics approach" necessary for developing an ethically accountable AI is based on the evidence that "neither people, nor the environment, are static; what society deems fair and ethical changes overtime. The concept of fairness and ethical practice is, therefore, a moving target and not something that can have a final answer or can be 'solved' once and for all" ([35], p. 2).

It is, therefore, worth emphasizing that such a dynamic approach to relational ethics is what also characterizes Amartya Sen's revolutionary theory of the *comparative distributive justice* in the realm of social and economic theories [36], for which he was awarded with the Nobel Prize in 1998, and which is modeled in the formal framework of the so-called "social choice theory" [37]. Now, as we discussed elsewhere [2], the fundamental computational challenge impeding the extensive use of Sen's theory in social and economic modeling is precisely the same as that addressed in the above quoted NISP paper, namely the necessity of a *dynamic* continuously updated weighing of the variables that in principle in relational ethics, like in the case of data streaming [22,26], requires a coalgebraic *dynamic* and not *statistical* approach to machine learning in AI. The classical statistical approach, indeed, is thought of as dealing with "big" but "static" bases of data. Therefore, it is not casual, but scientifically relevant, that in the common framework of the operator algebra approach, and then of the Category Theory logic, the same *coalgebraic* approach for formalizing Kripke's modal relational semantics (that is evidently also the (deontic) logic of the relational ethics) is the same as that used for modeling quantum dissipative systems in the condensed matter physics of dissipative biological and even neural systems, both the natural and the artificial ones (see note 3). The future of neuroethics and of an ethically accountable AI must evidently also go along this coalgebraic path.

Funding: This research received no external funding.

References

1. Basti, G. The Post-Modern Transcendental of Language in Science and Philosophy. In *Epistemology and Transformation of Knowledge in in Global Age*; Delic, Z., Ed.; InTech: London, UK, 2017; pp. 35–62.
2. Basti, G.; Capolupo, A.; Vitiello, G. The Computational Challenge of Amartya Sen's Social Choice Theory in Formal Philosophy. In *The Logic of Social Practices. Studies in Applied Philosophy, Epistemology and Rational Ethics 52*; Giovagnoli, R., Lowe, R., Eds.; Springer Nature: Berlin, Germany; New York, NY, USA, 2020; pp. 87–119.
3. Gazzaniga, M.S. *Who Is in Charge? Free Will and the Science of the Brain*; Harper Collins Publ.: New York, NY, USA, 2011.
4. Putnam, H. Minds and Machines. In *Dimensions of Mind*; Hook, S., Ed.; New York UP: New York, NY, USA, 1960; pp. 362–385.

$A \rightarrow A \times A$ of coalgebras for modelling the energy balance (summation) system-environment: $E_{sys} - E_{env} = 0$, characterizing dissipative systems in far-from-equilibrium conditions (i.e., with $T > 0$). Similarly, a coalgebraic semantics is necessary for modelling Kripke's "modal relational semantics" because Kripke's "possible worlds" are as such "worlds-in-relationship" with (an environment of) other worlds. Differently, for instance, from the "possible worlds" of Carnap's modal logic and of its "logical atomism" [1,2,31–34].

5. Pylyshyn, Z.W. *Computation and Cognition: Toward a Foundation for Cognitive Science*; MIT Press: Cambridge, MA, USA, 1986.
6. Fodor, J.A. *Modularity of Mind: An Essay on Faculty Psychology*; MIT Press: Cambridge, MA, USA, 1983.
7. Fodor, J.A. Methodological Solipsism Considered as a Research Strategy in Cognitive Science. *Behav. Brain Sci.* **1980**, *3*, 63–73.
8. Damasio, A. *Descartes' Error: Emotion, Reason, and the Human Brain*; Putnam Publishing: New York, NY, USA, 1994.
9. Kant, I. *Grundig for the Metaphysics of Morals*; Hackett: London, UK, 1993.
10. Damasio, A. *Self Comes to Mind: Constructing the Conscious Brain*, 1st ed.; Heinemann: London, UK, 2010.
11. Damasio, A. *The Strange Order of Things. Life, Feeling and the Making of Cultures*; Pantheon Books: New York, NY, USA, 2018.
12. Libet, B.; A Gleason, C.; Wright, E.W.; Pearl, D.K. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* **1983**, *106*, 623–642.
13. Haynes, J.-D.; Sakai, K.; Rees, G.; Gilbert, S.J.; Frith, U.; Passingham, R.E. Reading Hidden Intentions in the Human Brain. *Curr. Boil.* **2007**, *17*, 323–328.
14. Basti, G. *Filosofia Dell'uomo*, 2nd ed.; ESD: Bologna, Italy, 2003.
15. Levy, N. *Neuroethics: Challenges for the 21st Century*; Cambridge UP: Cambridge, UK, 2007.
16. Levy, N. *Consciousness and Moral Responsibility*; Kindle edition; Oxford UP: Oxford, UK, 2014.
17. Freeman, W.J. A proposed name for aperiodic brain activity: stochastic chaos. *Neural Networks* **2000**, *13*, 11–13.
18. Freeman, W.J. *How Brains Make up Their Minds*; Columbia UP: New York, NY, USA, 2001.
19. Freeman, W.J.; Vitiello, G. Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics. *Phys. Life Rev.* **2006**, *3*, 93–118.
20. Freeman, W.J.; Vitiello, G. Dissipation and spontaneous symmetry breaking in brain dynamics. *J. Phys. A Math. Theor.* **2008**, *41*, 304042.
21. Basti, G.; Capolupo, A.; Vitiello, G. Quantum field theory and coalgebraic logic in theoretical computer science. *Prog. Biophys. Mol. Boil.* **2017**, *130*, 39–52.
22. Basti, G. The quantum field theory (QFT) dual paradigm in fundamental physics and the semantic information content and measure in cognitive sciences. In *Representation and Reality in Humans, Other Living Organisms, and Intelligent Machine*; Dodig-Crnkovic, G., Giovagnoli, R., Eds.; Springer: Berlin, Germnay; New York, NY, USA, 2017; pp. 177–210.
23. Basti, G. Intelligence and reference. Formal ontology of the natural computation. In *Computing Nature. Turing Centenary Perspective*; Dodig-Crnkovic, G., Giovagnoli, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 139–159.
24. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
25. Blasone, M.; Jizba, P.; Vitiello, G. *Quantum Field Theory and its Macroscopic Manifestations, in Boson Condensation, Ordered Patterns and Topological Defects*; Imperial College Press: London, UK, 2011.
26. Rutten, J.J.M. Universal coalgebra: a theory of systems. *Theor. Comput. Sci.* **2000**, *249*, 3–80.
27. Frölich, H. (Ed.) *Biological Coherence and Response to External Stimuli*; Springer: Berlin, Germany, 1988.
28. del Giudice, E.; Doglia, S.; Milani, M.; Vitiello, G. A quantum field theoretical approach to the collective behavior of biological systems. *Nucl. Phys.* **1985**, *B251*, 375.
29. del Giudice, E.; Pulselli, R.; Tiezzi, E. Thermodynamics of irreversible processes and quantum field theory: An interplay for understanding of ecosystem dynamics. *Ecol. Model.* **2009**, *220*, 1874–1879.
30. Montagnier, L.; Aïssa, J.; Capolupo, A.; Craddock, T.; Kurian, P.; Lavallee, C.; Polcari, A.; Romano, P.; Tedeschi, A.; Vitiello, G. Water Bridging Dynamics of Polymerase Chain Reaction in the Gauge Theory Paradigm of Quantum Fields. *Water* **2017**, *9*, 339.
31. Kupke, C.; Kurz, A.; Venema, Y. Stone coalgebras. *Theor. Comput. Sci.* **2004**, *327*, 109–134.
32. Blackburn, P.; de Rijke, M.; Venema, Y. *Modal logic. Cambridge Tracts in Theoretical Computer Science*; Cambridge UP: Cambridge, UK, 2002.
33. Venema, Y. Algebras and co-algebras. In *Handbook of Modal Logic*; Blackburn, P., van Benthem, F.J.F., Wolter, F., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 331–426.
34. Goranko, V.; Otto, M. Model theory of modal logic. In *Handbook of Modal Logic*; Blackburn, P.F., van Benthem, J.F., Wolter, F., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 252–331.

35. Birhane, A.; Cummins, F. Algorithmic Injustice: Toward a Relational Ethics. 16 December 2019. Available online: https://arxiv.org/pdf/1912.07376v1.pdf (accessed on 9 April 2020).
36. Sen, A.K. *Collective Choice and Social Welfare*; Expanded Edition; Kindle Edition; Penguin Ltd.: London, UK, 2017.
37. Endriss, U. Logic and social choice theory. In *Logic and Philosophy Today*; Gupta, A., van Benthem, J., Eds.; College Pubblications: London, UK, 2011; pp. 333–377.