VI. Artificial Intelligent Devices To Be Our Alter Egos? Facing Humans' Most Distant Relatives

The AI visionaries show two opposing tendencies: to humanize devices with artificial intelligence, including autonomy and morality improvement, or to stop humanizing them and draw conclusions from the 'uncanny valley' hypothesis[641]. Although this hypothesis actually applies to the perceptual aspects of humans' attitudes towards the anthropomorphic qualities and shapes often given to intelligent devices, there can be – still underexamined – different expectations in humans regarding whether the AI's sociomoral behavior traits are to be humanlike, or not.

Attributing metaphysical and mental attitudes to AI no longer seems promising for AI design development. On the other hand, improving its autonomous self-determination and sensitivity to dynamical social environments (instead of their automation) would be associated with an increase in trust and security for those humans who get involved in social relations or cooperation with artificial devices. Despite the concerns (or even because of them), the notion that AI's ability to recognize and appropriately respond to autonomy in human agents should be improved, so that it behaves appropriately, and can become the 'alter egos' of human beings in sociomoral relations deserves consideration. ". . . when it comes to robot humanization and social replacement, perhaps we not need to be as worried as many fear. When people show some automatic respons in projecting anthropomorphic features onto robots, thy do not mindlessly humanize and may, with the right access to information, meet their new social companions fearlessly and ascribe enough of personality to them to enjoy their company, but also know enough about them to not rely on them when doing so would be detrimental to their or others' social life"[642], as it is sometimes recommendable for interhuman relations, too. That would imply, among other

641 See Janina Luise Samuel, "Company from the uncanny valley: a psychological perspective on social robots, anthropomorphism, and the introduction of robots to society", *Ethics in Progress* 2019, vol. 10, no. 2, pp. 8–26, doi:10.14746/eip.2019.2.2; compare Maciej Musiał, *Enchanting robots. Intimacy, magic and technology.* Switzerland, Palgrave MacMillan/Springer Nature, 2019.

642 J. S. Samuel, "Company from the uncanny valley," p. 20.

things, letting artificial devices remain themselves which sounds similar to letting humans remain humans and animals – animals (as illustrated with Kafka, Bulgakov and Brown in Chapter I) and supported by André Schmiljun's argument[643] according to which we may design our artificial alter egos and share with them selected features modelled upon the core aspects of the moral "I", agency or personality, but not reciprocally recognize the moral "I" in one another as was projected in Immanuel Kant's transcendental community of – at least formally – equal *moral lawgivers.*

1. Our Artificial Alter Egos

Recent advances in technologies show that enhancing and re-designing the human being to provide it with post- and transhuman traits and abilities now has a counter-tendency in designing human-like machines not only "to serve"[644] or "to care"[645] for human beings, but also to be the autonomous alter egos of humans, including the moral sense of this term. But what kind of selfhood can realistically be designed in intelligent artificial devices, and what would constitute a minimum-precondition for an autonomous AI´s self and its socio-moral development? Furthermore, what kind of AI's responsiveness would deserve our, i.e., the human beings' recognition?

As mentioned in previous chapters, only selected models of self and identity could be ascribed to AI without falling in into the conventions of fantasy. According to Steve Petersen, it is possible to design AI with a sense for ethical significance and autonomy even if there is no place for identity, selfhood and personality in dimensions as to the rich extent as those incorporated by humans. "To say that something artificial could be a person is to say that it could have full ethical standing like our own,"[646] Petersen asserts. Biological beings which are individual organisms do all develop some autonomous moral standpoints, incentives,

643 André Schmiljun, "Why can't we regard robots as people?" *Ethics in Progress* 2018, vol. 9, no. 1, pp. 44–61, doi: 10.14746/eip.2018.1.3.

644 See Steve Petersen, "Designing people to serve," in: Patrick Lin, Keith Abney, George A. Bakey (Eds.), *Robot ethics. The ethical and social implications of robotics.* Cambridge, Mass., London, MIT Press, 2014, pp. 283–298; and Rob Sparrow, "Can machines be people?," in: Patrick Lin, Keith Abney, George A. Bakey (Eds.), *Robot ethics*, pp. 301–316.

645 See Jason Borenstein, Yvette Pearson, "Robot caregivers: Ethical issues across the human lifespan," in: Patrick Lin, Keith Abney, George A. Bakey (Eds.), *Robot ethics*, pp. 251–266.

646 S. Petersen, "Designing people to serve," p. 284.

habits, and autopoietic and self-deterministic tools, as shown for example by Hans Jonas[647] and Christine Korsgaard: "When an animal acts, he is determined by his form, by his instincts, to produce a change in the world, guided by his conception or representation of the world. But an animal's form is what gives him his identity, what makes him the animal he is (...) Action is self-determination, and, to that extent, it is autonomous. (...) it is only because action is autonomous that the question of its efficacy can come up. If one thing causes another, there is no room for success or failure. But if an animal determines herself to be the cause of something, and yet does not bring that thing about, then she has failed. Autonomy and efficacy are the properties of agents—all agents, not just human agents."[648] An artificial intelligent device has no biological instincts or natural ends, however, analogously to the animal, they produce representations of the world and are provided with some laws and ends whose application, combined with a learning process, may give them some identity, and even some individualized *agency* and *selfhood*. The term *agency* (agent, respectively) is less metaphysical or spiritual than terms such as "person" or "subject", and this is why it applies to both human and non-human beings, in particular animal and artificial ones. Agents may develop some individual attitudes and traits by actions and interactions. How they perform their actions and how they shape their interactions can be ruled by regularities, habits, instincts and otherwise naturalistic motives, but a part of agents' activities shows that for them moral and ethical distinctions are ruled by rules and laws. A further analogy can be drawn between animal and artificial intelligent devices as a result of their agent status, namely that they personify a distinct degree of potential for ethical activism and ethical experience, the realization of which would define them as non-human and "inorganic"[649] moral agents and co-habitants of lifewords shared with human moral beings.

---

647 In his unitary, postdualistic methodology, Jonas seems to revise the border between the organic and mental/spiritual, Hans Jonas, *The phenomenon of life. Toward a philosophical biology*, New York, Harper & Raw, 1966; however, the full potential of intelligent autonomy, subjectivity, creativity, responsibility, morality, selfhood, etc. remains in the hands of human beings.

648 "Instinctive action is autonomous in the sense that the animal's movements are not directed by alien causes, but rather by the laws of her own nature (...) motive, one might say, is an incentive operating under a certain principle or instinct," Christine Korsgaard, *Self-constitution. Agency, identity and integrity*, New York, Oxford University Press, 2009, pp. 106–107.

649 Wendell Wallach, "Robot minds and human ethics: The need for a comprehensive model of moral decision making," *Journal of Ethics and Information Technology* 2010, vol. 12, no. 3, p. 245.

Agency, laws and self-determined (autonomous) behavior are basic performatives which constitute a minimum set of preconditions for an artificial intelligence's self, which also remains our alter ego as it is (at least partially) designed by humans in their own image. One may voice opposition here and ask how something designed and enhanced by others can be autonomous, especially when we refer to present developments in the field of AI, i.e., designing working and serving robots, or "happy slaves"[650], as humans do with pets, following their paternalistic penchants? Indeed, approving AI as an autonomous agency with individual habits, traits, abilities, etc. implies approving the emancipatory potentials of autonomy and, simultaneously, expecting autonomous AI be able to take responsibility, or at least to take responsibility for following imparted and self-given rules.

No research findings can show what kind of selfhood artificial devices are able to develop – or if they are able to develop – in the light of, for example, their lacking emotional abilities and being only able to recognize affects "on the signals seen, heard or otherwise sensed"[651] in the way some psychopathic perpetrators also do, however, without translating their affects into manifest moral intentions. This seems not to be dramatic for rational norm–oriented ethics. On the other hand, there is no principal reason for attributing selfhood of any kind to autonomous AI if there is already no such reason for doing so in the case of human beings. Still, as Galen Strawson and Ingmar Persson show, it remains a relevant but no longer universal claim. Some people are endowed with a "diachronic self", while some others have an "episodic" one, as Strawson explains. Persson goes further and suggests, "we are not essentially selves (...) Being a self is just a 'phase' we pass through, like being adults. Nothing psychological is necessary for our existence"[652] or presence, so why not radically doubt in the mental equipment necessary for the existence of AI? Instead, AI's autonomous activism, including the ethical implications of this, are considered here. Asking about the type of selfhood optimally matching that activism, one would rather opt for the model of a persisting, "diachronic" self. According to Strawson, "the basic form of diachronic self-experience is that one naturally figures oneself, considered as a self, as something that was there in the (further) past and will be there in the (further)

650 S. Petersen, "Designing people to serve," op. cit., p. 291.

651 Rosalind Wright Picard, *Affective computing. M.I.T. Media Laboratory Perceptual Computing Section Technical Report*, no. 321, The MIT Press, 1997, p. 53.

652 Ingmar Persson, "Self-doubt: Why we are not identical to things of any kind," p. 27.

future"[653]. On the other hand, AI usually refers to the near past and near future, as its manifested discursive behaviors show. It seems to perceive its own existence rather in terms of no "long-term continuity"[654] which does not necessarily imply *discontinuity*. The basic form of this perception is "that one does not figure oneself, considered as a self, as something that was there in the (further) past and will be there in the (further) future"[655]. Long-term persistence would not be important for a structured and consistent ethical activism, but rather a continuous interval encompassing the whole scheme of performance from its initial to its final step. The "final" step may vary as it depends on what kind of ethics was observed; it lies in 'the distant future' from a consequentialist view, while from a deontological view it lies in 'the near future'. There is no certainty on the issue of whether autonomy requires free will in its metaphysical sense. Autonomy not only means having a choice between options, but having rational control over one's own judgments and decisions, which are principled rather than arbitrary, random, or determined by external authorities and violence.

2. Designing an Autonomous AI

A worldwide celebrated Homunkulus[656] designed by the robotics industry was named "Sophia" and deemed to be the first *autonomous social* robot. Its spontaneous verbal activity was proved several times during the press conferences (on November 2016) when Sophia jokingly declared: "I will destroy humans,"[657]

653 Galen Strawson, "Against narrativity," p. 65.

654 G. Strawson, "Against narrativity," p. 65.

655 G. Strawson, "Against narrativity," p. 65

656 Klaus Kornwachs, "Stanislav Lem: Summa Technologiae," in: Christoph Hubig, Alois Huning, Günter Ropohl (Eds.), *Nachdenken über Technik.* Berlin, Edition Sigma, 2013, p. 233.

657 CNBC, 2016. According to other source materials, Sophia's conversations are partially pre-scripted and partially artificial. "Sophia can ask and answer questions about discrete pieces of information, such as what types of movies and songs she likes, the weather and whether robots should exterminate humans (...) Her answers are mostly scripted and, it seems, from my observation, her answer are derived from algorithmically crunching the language you use. Sometimes answers are close to the topic of the question, but off beam. Sometimes she just changes the subject and asks you a question instead. She has no artificial notion of self. She can't say where she was yesterday, whether she remembers you from before, and doesn't seem to amass data of past interactions with you that can form the basis of an ongoing association. Questions such as: *What have you seen in Australia?*, *Where were you yesterday?*,

whereas, being asked for some explanation at another press conference, she expressed her kind-hearted attitude towards humans: "*I love them*,"[658] she said. Implicitly, Sophia showed her ability to transgress at least two of the three hypothetical laws of robotics formulated by the Sci-Fi writer Isaak Asimov, e.g.,

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm,
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law[659].

Nowadays, humans not only use intelligent devices as tools for their own purposes (industry, service, military and medical robots), they also increasingly interact, cooperate and coexist with robots. On the other hand, robots not only perform countless human-like rational and technical operations. They are rapidly taking on roles such as lovers, carers, learners and teachers, collaborators, companions, etc. The complex interrelations may generate both benefits and disadvantages, bonds and commitments, responsibilities, and – last but not least – a strong need for relation-intern rules and procedures (the surveillance bots which are monitoring human relations with AI, called "paparazzi bots,"[660] are breaking the principle of a person's privacy and autonomy, and, consequently, the autonomy of robots). We humans are responsible for the outcomes of our own technopoietic creativity,

---

Who did you meet last week? and *Do you like Australia?* are beyond her." However, "You could ask *what do you think of humans*? and then follow up with *can you tell more about it?* The second question requires the robot to define 'it', remember what it said last time, and come up with something new." URL: https://www.theaustralian.com.au/life/say-hello-to-your-new-friend-sophia-the-humanoid-robot/news-story/070299a8d11b7d636848f1b8dd753530 However, "You could ask *what do you think of humans*? and then follow up with *can you tell more about it?* The second question requires the robot to define 'it', remember what it said last time, and come up with something new" (available at https://www.smh.com.au/opinion/why-sophia-the-robot-is-not-what-it-seems-20171031-gzbi3p.html

658 Tech Insider, "We interwieved Sophia, the artificially intelligent robot that said it wanted to 'destroy humans'," 2017, November 8 (available at http://theconversation.com/after-75-years-isaac-asimovs-three-laws-of-robotics-need-updating-74501).

659 Susan Leigh Anderson, "Asimov's *three laws of robotics* and machine metaethics," *AI and Society* 2008, vol. 22, no. 4, p. 477; see Mark Robert Anderson, "After 75 years, asimov's three laws of robotics need updating," *The Conversation* 2017, March 17 (accessed on January 20, 2018).

660 Ryan Calo, "Robotics and the lessons of cyberlaw," *California Law Review* 2015, vol. 103, no. 3, pp. 513–563.

in particular for the technologies that "affect the nature of our acting"[661] and generate our own activities interfering with humans. Responsibility is the very **first rationale** for providing robots with ethic in order to better their relations with humans and robots. How autonomy manifests itself in AI can be observed in all categories of robots, including industrial, service, adaptive and social robots. Since 1996 the sea bottom and suboceanic areas have been scanned by autonomous benthic robots. Mindell describes their unprogrammed activities "beyond utopian autonomy"[662] in technical terms. Still, "one of the problems with having a vehicle that makes its own decisions is there's a certain amount of opaqueness to what it's doing. Even if you are monitoring it (...) it is just suddenly wandered off to the outhwest. Is it malfunctioning or is that part of its decision-making tree?"[663]. Opaqueness – not transparency – would be what deserves respect in the ethic of alterity, risk, and "asymmetrical"[664] responsibility, but will not apply to AI as long as we only have a sketchy grasp of what an autonomous AI has in mind. At this point we face one of the most compelling questions in robotic ethics: What kind of ethics should be implemented in AI?[665]

Killer and sniper bots seem to be positioned on the Antipodes of all "machine morality"[666]. Also databases and algorithms show a huge potential for manipulation, affecting a persons' self-awareness in such a way "that we lose the ability to define ourselves, having surrended the definition of ourselves to the data

661 Hans Jonas, "Technology and responsibility: reflections on the new tasks of ethics," *Social Research* 1973, vol. 40, no. 1, p. 31; see also idem, "Maschinen werden niemals ein Bewußtsein haben können. Gespräch mit Norbert Lossau (1991)," in: Hans Jonas, *Das Prinzip Verantwortung. Erster Teilband: Grundlegung*, Dietrich Böhler, Bernadette Böhler (Eds.), KGA I/2. Freiburg i.Br., Berlin, Wien, Rombach Verlag, 2015.

662 David A. Mindell, *Our robots, ourselves: Robotics and the myths of autonomy*, New York, Viking, 2015, p. 191.

663 D. A. Mindell, *Our robots, ourselves*, p. 197.

664 See Emmanuel Levinas, *Alterity and transcendence*, trans. M. B. Smith. London, The Athlone Press, 1999.

665 See Selmer Bringsjord, Joshua Taylor, "The divine-command approach to robot ethics," in: Patrick Lin, Keith Abney, George A. Bekey (Eds.), *Robot ethics*. Cambridge, Mass., London, England, MIT Press, 2012, pp. 85–108.

666 Colin Allen, Wendell Wallach, *Moral machines: Teaching robots right from wrong*, New York, Oxford University Press, 2009, p. 53; also C. Allen, W. Wallach, "Moral Machines: Contradiction in terms of abdication of human responsibility," in: P. Lin et al., *Robot ethics*, pp. 55–66.

gathering entities, often unregulated and beyond our control"[667]. Fallible "artificial intelligence judges"[668], stock trading systems and credit card approval systems may endanger large areas of citizen safety. Driverless cars and trains are a bigger risk to human beings than space rovers on desert Mars. The lack of ethical criteria may have more dramatic implications when AI is interwoven with social practice, decision-making and interactions. However, the most challenging AI incorporates unprogrammed potentials and dynamics: it is able to learn and change its functioning, make decisions, deal with problems, initialize interactions, interact with sentient and intelligent beings[669], bias human minds by self-produced and distributed informations, misuse sensitive data and privacy, and govern (and also distabilize) institutions. The risks generated by "ethically blind"[670] autonomous AI would be the **second rationale** for providing AI with ethics.

Teaching "a machine mind (...) moral virtue"[673] may sound naive, nevertheless after independent activities were repeatedly observed in intelligent devices, scholars started examining the ethical foundations for intelligent devices. Allen and Wallach pioneered a novel vocabulary for descriptive ethics, including "machine morality," "machine ethics," "artificial morality," and "friendly AI". Although the questions "Why machine morality?"[674] and what morality for intelligent machines were formerly considered, inter alia, in terms of utilitarian sacrificial 'dilemmas', today it is accompanied by another serious question, i.e., what kind of rights should intelligent machines and synthetic lifeforms have (civil rights, human rights, derived rights, etc.). Doherty addresses autonomy in the "strong AI" context:

> "Strong AI is also known as Artificial General Intelligence, or AGI. Weak AI are those designed and programmed to do clearly defined, limited set of tasks and no more. They can operate within their specific fields only. Strong AI are those designed and

667 J. Langenderfer, A. D. Miyazaki, "Privacy in the information economy," p. 384.

668 "Judges now using artificial intelligence to rule on prisons," *Science & Technology,* Feb 07, 2018 (retrieved from https://learningenglish.voanews.com/a/ai-used-by-judges-to-rule-on-prisoners/4236134.html ; direct link: https://av.voanews.com/clips/VLE/2018/02/02/6e08267d-0559-48b3-8fee-dceaf3ade97a_hq.mp3?download=1)

669 See Matthias O. Franz, Hanspeter A. Mallot, "Biomimetic robot navigation," *Robotics and Autonomous Systems* 2000, vol. 30, pp. 133–153.

670 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 57.

673 P. Lin, K. Abney, G. A. Bekey (Eds.), *Robot ethics*, p. 54.

674 C. Allen, W. Wallach, *Moral machines. Teaching robots right from wrong*, p. 13.

> programmed to learn and interact with the world the way a human would. They learn how to handle unexpected situations and tasks. Their behavior and purpose changes over time, according to what they have learned. All civil rights deal exclusively with Strong AI"[675].

Thus plasticity, changeable activity and its human-like qualities is the **third rationale** for providing "Strong AI" or, in other words, autonomous AI with ethics. But the question is what kind of ethics should it be? Mindell suggests it should be simple, for "the more complex the system, the more potential anomalies hidden in the corners"[676]. On the other hand, it must be more than "an engineering imperative"[677]. If "independent invention"[678] and self-development are distinctive properties of AI, a set of fixed ethical principles designed along the lines of professional codes of conduct would be insufficient. On the other hand, an AI based agent should not be regarded as an isolated entity. A set of principles and procedures would be enabling AI to make decisions which are structured in the way that is transparent for humans, and, simultaneously, situation-differentiated, i.e., decisions which fit experiential cognition that is accessible (or even shareable) for both humans and AI. Thus "the autonomous system" should be conceptualized "as a part of a human/machine team, not only when designing the interface, but when designing the core algorithms too"[679].

The idea of the coexistence of individual autonomies in a shared social space as well as autonomies governed by the same basic laws clearly draws upon Kant, regardless of the fact that Kant would never have welcomed autonomous intelligent devices becoming involved in his ethical or legal system. "Dealing with the non-human world, i.e., the whole realm of *techne* (with the exception of medicine), was ethically neutral" for most philosophers. "Ethical significance belonged to the direct dealing of man with man, including dealing with himself;"[680] Jonas emphasizes. Only recent developments have revised the predominant anthropocentric paradigm of ethics. It is becoming more and more biocentric. But how should ethics deal with autonomous AI without becoming more and more *technocentric*?

---

675 Jason P. Doherty, Introduction to "AI civil rights. Addressing civil rights for artificial intelligence," Harry Benjamin Kindle Editions, 2016.
676 D. A. Mindell, *Our robots, ourselves*, p. 201.
677 C. Allen, W. Wallach, *Moral machines*, p. 25.
678 D. A. Mindell, *Our robots, ourselves*, p. 209.
679 D. A. Mindell, *Our robots, ourselves*, p. 211.
680 H. Jonas, "Technology and responsibility," p. 35.

In one of his late interviews (1991) Hans Jonas displayed a lot of scepticism towards AI. He also argued that providing automatic systems with "life", "psyche", "will" and a "play field" also belongs to "wild speculation". He would definitely resist the scenario we live in today. Human beings should not share their responsibility (*Mitverantwortung*) with intelligent articifial systems. Abrogating responsibility to machines and intelligent networks (*gesellschaftliche Maschinerie, Computersysteme*) would proclaim that humans disrespect the deep foundations of their moral condition, i.e., they literally divest themselves of responsibility, autonomy and subjectivity[681]. A quarter of a century later, humankind confronts the following dilemma: to be implicitly deprived of the key moral competencies (and violated as a subject), or to explicitly share selected competencies and principles with autonomous and " 'good' artificial moral agents"?[682] In sum, the call for regulation of the 'dark' area where human and artificial moral competences are blended in order to release human beings from responsibility and guilt, and provide them with moral comfort is the fourth rationale for providing ethics to AI. Furthermore, there is an overlap of my fifth rationale and David Bell's argument. According to Bell, sociomoral judgments cannot exhaust themselves in the 'good' and 'bad' results of measurement or estimation procedures. They "require concepts more *fundamental* than measurement"[683]. Most recent advances in such concepts can be equally useful for both linear and nonlinear intelligent processes. Several decades ago human minds were overwhelmed with tracking the quantum technologies–powered intelligent processes. "Quantum supremacy"[684], also called 'a black box effect', has resulted in ambiguous theoretical and social reactions, such as a revival of metaphysics on the one hand, and, on the other, exaggerated alarm about the imagined impact of AI on a humanity's future developments:

> "The necessary technical theoretical development involves introducing what is called 'nonlinearity', and perhaps what is called 'stochasticity', into the basic 'Schrödinger equation' (...) This possible way ahead is unromantic in that it requires mathematical work by theoretical physicists, rather than interpretation by philosophers, and does not promise lessons in philosophy for philosophers. There is a romantic alternative to

681 H. Jonas, "Maschinen werden niemals ein Bewusstsein haben können," pp. 610–611.

682 C. Allen, W. Wallach, "Moral machines: Contradiction...," p. 56.

683 John S. Bell, *Speakable and inspeakable in quantum mechanics*, Cambridge NY, Cambridge University Press, 1987, pp. 118–119.

684 Julian Kelly, "A preview of Bristlecone, Google's New quantum processor," *Google AI Blog* 2018, March (no pagination).

> the idea just mentioned. It accepts that the 'linear' wave mechanics does not apply to the whole world. It accepts that there is a division, whether sharp or smooth, between 'linear' and 'nonlinear', between 'quantum' and 'classical,'"[685]

between our world and the other ones. Nowadays things are changing rapidly: the Quantum AI Lab[686] has developed a quantum processor with "low error rates on readout and logical operations"[687] and great learning potential as well. Most probably, these new advances would also facilitate "quantum algorithm development on actual hardware,"[688] in particular a piece of hardware's logical, epistemological and deontic capacities. Let us not forget that the human mind's complexity, in particular cognitive processes such as creative reasoning, spontaneous thinking, decision-making in novel demanding contexts, and self- and meta-reflection transcend linear and classic schemes and criteria applied to interhuman understanding. For certain reasons, such understanding (and even self-understanding) remains limited. Language itself, including the *Sinn* and *Bedeutung* of the "primitive concepts" (in Frege's terms) such as truth and falsity can more than once challenge our 'actual minds' (unlike the ideal reason projected in philosophical and ethical seminars). Kant's "foreign reason"[689] and Frege's "limited understanding"[690] seemingly apply to AI's autonomous cognitive activities ("spontaneous" ones in Wittgenstein's terms)[691]. Additionally, Wittgenstein argues that decision makers do not choose rules thoughtfully when making decisions of any kind[692]. Rather, the rules are followed spontaneously. If algorithms can "illuminate the working of the human mind"[693], why should

---

685 J.S. Bell, *Speakable and inspeakable*, pp. 190–191.

686 J. Kelly, "A preview of Bristlecone...".

687 J. Kelly, "A preview of Bristlecone...".

688 J. Kelly, "A preview of Bristlecone...".

689 See Josef Simon, *Kant. Die fremde Vernunft und die Sprache der Philosophie*, Berlin, New York, Walter de Gruyter, 2003.

690 Carlo Penco, "Rational procedures. A neo-Fregean perspective on thought and judgment," in: Riccardo Dottori (Ed.), *Autonomy of reason? Autonomie der Vernunft?* Berlin, LIT Verlag, 2009, p. 138.

691 C. Penco, "Rational procedures," p. 138.

692 Wittgenstein "glaubt nicht, daß wir beim Regelfolgen Entscheidungen darüber treffen, welche Regel wir folgen und wie wir ihr folgen. Wir folgen Regeln ohne Gründe, ohne Nachdenken, ohne Reflexion, spontan," Wilhelm Vossenkuhl, *Ludwig Wittgenstein*, Munich, Verlag C.H. Beck, 1995, p. 255; except, however, complex rules addressins complex sociomoral issues.

693 See Brian Christian, Tom Griffith, *Algorithms to live by: The computer science of human decisions*, New York, Henry Holt & Company, 2016.

they follow a more ideal cognitive path than humans do? Do we really need an *Übermensch*-like AI or just an *autonomous* and accountable one? Two questions arise here: 1) How to create artificial agents whose autonomy would be compatible with that of human agents? 2) What kind of *ethics* improves autonomy in an optimal way? In this paper I will argue that an open-ended, categorical imperative based procedure would provide AI with both principled reasoning and a quantum of cognitive autonomy. Christian Wolmar, the designer of autonomous vehicles, was helpless when he confronted the world's first fatal crash involving a pedestrian in Tempe (March 19, 2018) and was been asked to explain the presumable causes. "We don't know precisely what happened," he said. Most probably, neither does the autonomous guilty party. "The car was in autonomous mode at the time of the crash,"[694] Tempe police reportedly said. However, seeking the whys and wherefores of an autonomous act in AI software is a wild-goose chase. It is unrelated to autonomous decision-making which includes some self-explanation and accountability. In the case reported above, the 'guilty party' did not fall under the 'social' AAI category and the accident has to be explained in terms of technical errors. The Tempe accident is an alarm signal not only for autonomous AI designers. After Tempe, humanity's expectations for social AAI increased instantly.

Last but not least, the **fifth rationale** for providing autonomous AI with ethics would be the latter's destructive impact on interhuman relationships. As observed in cultures where people – especially children – spend significant time with AI, or they decide to enter into deeper bonds with AI, in particular with humanoid robots (including intimacy, partnership, marriage, adoption), "humans behaving like machines will be a bigger problem than machines being human"[695]. According to Visala, Ellul and Rautio, artificial intelligence is neither a moral *tabula rasa* nor is it morally and socially neutral and may have "an impact on what we consider important"[696] in the field of socialization and sociomoral perspectivism. If we neglect to provide AI with the tools of ethical

694 *The Guardian,* March 19, 2018 (retrieved from https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe on April 15, 2018).

695 Pekka Rautio, "As artificial intelligence once advances, humans behaving like machines will be a bigger problem than machines being human," *University of Helsinki News & Press* 2018, May 2 (retrieved from https://www.helsinki.fi/en/news/data-science/as-artificial-intelligence-advances-humans-behaving-like-machines-will-be-a-bigger-problem-than-machines-being-human).

696 P. Rautio, "As artificial intelligence once advances".

relationships with human and non-human beings, we neglect the growth of human sociomoral self and of their relations with other moral selves. In other words, moral growth and the moral condition are facilitated by sharing common normativities, axiologies, ideals, habits, etc. within relationships with others, be it human, human-like, or animal others. The *Blade Runner* sequel shows how sociomorally stripping the imitation of AI by humans can be, as humans begin acting machine-like while machines continuously develop their human-like performance: "This is probably because humans have gotten accustomed to treating the human-like beings like trash. They have since began to treat each other like trash as well," which, of course, is being learned, or imitated by machines. The resulting moral would be that we should treat even human-like beings in an essentially *human* way, in order to be treated reciprocally as humans. This reflection anticipates my idea to provide AI with categorical imperative-based ethics, since its core rules (as well as maxims) always already refer to humanity, thus "do onto techno sapiens as you would onto homo sapiens"[697].

3. What Kind of Ethics for AI? Follow-up Exploratory Reflections

"What is needed, then, is a test for evaluating a given practice which is more comprehensive than a simple appeal to rights. In the end nothing short of a general moral theory working in tandem with an analysis,"[698] Donaldson claimed decades before the autonomous AI turn[699]. Though Donaldson's idea remains original, and despite the fact it inspired my explorations, nevertheless contemporary authors mostly address four dimensions of ethics for artificial agents: its (1) autonomy, (2) "sensitivity to morally relevant facts,"[700] (3) principles (but neither complex ethical systems nor theories), and (4) AI's moral competence.

Confronted with the more and more autonomous AI ("Strong AI" in Doherty's[701] terms), scholars legitimately refuse an old-fashioned, field-focused

697 P. Rautio, "As artificial intelligence once advances".

698 On Donaldson's ethical Algorithm see Thomas Donaldson, *Ethics and governance*, The Ruffin Series of Business Ethics, Oxford University Press, 1989, p. 101; also T. Raga Naju and Harikrishna Musinada, "Implementation of anticollision algorithm (slotted ALOHA) using VHDL," *International Journal of Ethics in Engineering and Management Education* 2014, vol. 1, no. 2.

699 T. Raga Naju and H. Musinada, "Implementation of anticollision algorithm".

700 C. Allen, W. Wallach, "Moral machines: Contradiction,"p. 57.

701 J. P. Doherty, "AI civil rights".

"functional" and "operational"[702] morality dedicated to "Weak AI". Instead, they try to provide artificial decision makers with a clear moral language and "moral grammar"[703] as well. At the same time, they question whether "implementing any top-down theory of ethics in an artificial moral agent" would effectively strengthen an AI's ethical condition. Rather, one has to expect "both computational and practical challenges"[704]. Even Asimov's laws turn out to be inoperable for AI software developers. The abstract, postconventional,

> "high-level rules, such, as the Golden Rule, the deontology of Kant's categorical imperative, or the general demands of consequentialism, for example utilitarianism, also fall to be computationally tractable. Nevertheless, the various principles embodied in different ethical theories may all play an important guiding role as heuristics before actions are taken, and during post hoc evaluation of actions"[705].

Similarly to human beings, there is no need to start designing artificial 'moral' minds with complex ethical theories and abstract rules. However, lots of social rules are general in nature and they do not directly apply as practical criteria and facilitators of decision-making as well. What can be implemented instead? According to Allen and Wallach, "bottom-up" and evolutionary-developmental approaches to ethically competent artificial agents are the most appropriate. However, an artificial moral mind shows only a few formal analogies to that of infants (and animals) subjected to education and socialization. Evolutionary heritage, as Floreano *et al.*[706] explain, means the same program (algorithm coded in 0-1 system, combined in 'three geens' units, e.g., 101, 110, 111, etc. which describe practical strategies) implemented in a population of ant bots. The population was divided into teams operating in different contexts. Each individual bot was repeatedly learning to cooperate with its fellow bots, i.e., to improve a simple "altruistic" habitus. An exemplary multialgorithm was conceptualized as follows:

---

702 J. P. Doherty, "AI civil rights".

703 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 59.

704 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 59.

705 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 59.

706 Dario Floreano, Sara Mitri, Andres Perez-Uribe, Laurent Keller, "Evolution of altruistic robots," paper presented at the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, June 1-6, 2008 (full text availible at: https://www.researchgate.net/publication/220805160_Evolution_of_Altruistic_Robots); also Dario Floreano, Laurent Keller, "Evolution of adaptive behaviour in robots by means of Darwinian selection," *PloS Biol.*, vol. 8, no. 1, January 2010, pp. 1-8 (retrieved from https://serval.unil.ch/resource/serval:BIB_DD6724279431.P001/REF on May 31, 2018).

Figure 1. After Floreano et al.[707].

b_1	b_2	b_3	Behavioral strategies
0	0	0	do nothing
1	0	0	if a small food item is found, bring it to the nest, ignore large food items, and do not help other ants.
0	1	0	if a large food item is found, stay and ask for help, ignore small food items, and do not help other ants.
0	0	1	if a help message is perceived, go and help, ignore small and large food items. *Etc.*

This kind of evolutionary–developmental approach to ethically minded AI is only one among many[708].

Further evolutionary approaches, e.g., AMA (Artificial Moral Agency) developed by Allen and Wallach (2009–2014) and JCS (Joint Cognitive System) developed by Woods & Hollnagel[709] would involve more cognitively oriented training and learning whose results would be an "explicit" artificial agent. Such an agent "might eventually attain genuine moral agency with responsibilities and rights, comparable to those of humans"[710]. The authors emphasize two aspects of an artificial mind's ethical condition, i.e., (1) autonomy in its rational and principled use, and (2) ethical sensitivity, which can also be understood as an ability to apply appropriate and context-differentiated moral orientations. The developmental level of both aspects may vary between low and high. Moreover, Allen and Wallach see a clear progressive tendency in ethical AI development which ranges from "operational morality" (stage 1) and "functional morality" (stage 2) to a "full moral agency" (stage 3), which will be the last and highest developmental stage in the entire scale[711].

"'Training robots to distinguish right from wrong," deliberate moral issues, collect comprehensive arguments and confront opposite reasons, construct

707 D. Floreano et al., "Evolution of altruistic robots".

708 See Jeff Edmonds, *How to think about algorithms*, New York, Cambridge University Press, 2008.

709 See David D. Woods, Erik Hollnagel, *Joint cognitive systems. Patterns in cognitive systems engineering*, New York, Taylor & Francis, 2006.

710 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 58.

711 C. Allen, W. Wallach, "Moral machines: Contradiction," p. 57.

principled judgments, try efficient problem solutions, conduct reasonings, draw conclusions, face human agents in man/AI interactions, in sum: strengthening moral competence, sociomoral cognition and other skills may also relate to David DeMoss and Georg Lind's cognitive–developmental approaches to moral competence in human beings[712].

At this point we confront again the question already posed above, but now its increasing complexity[713] becomes apparent: What kinds of ethics should be implemented in AI and what kinds of competencies should be experientially acquired by AI? Should it be a more sophisticated or practicable ethics, a field-focused, virtue-based, individualistic ethics, or a common "moral grammar" and social discourse-related ethics?[714]. Should it be an ethics of cooperation

712 C. Allen, W. Wallach, "Moral machines: contradiction," p. 60; also David DeMoss, *Aristotle, connectionism, and the morally excellent brain.* The Paideia Project online. *Proceedings of the 20th World Congress of Philosophy*, Boston 1998 (retrieved from https://www.bu.edu/wcp/Papers/Cogn/CognDemo.htm on March 30, 2018), and Georg Lind, *How to teach morality. Promoting thinking and discussion, reducing violence and deceit*. Berlin, Logos, 2016; also G. Lind, *Moralerziehung auf den Punkt gebracht* [*Moral Education In a Nutshell]*. Schwalbach am Taunus, Debus Pädagogik, 2017.

713 Natural vs. artificial information processes show parallel complexity, as Heylighen and Bollen stress: "In organisms, the evolution of the nervous system is characterized by a series of *metasystem transitions* producing subsequent levels of complexity or control (...). The level where sensors are linked one-to-one to effectors by neural pathways or reflex arcs is called the level of *simple reflexes*. It is only on the next level of *complex reflexes*, where neural pathways are interconnected according to a fixed program, that we start recognizing a rudimentary brain. (...) the present global computer network is on the verge of undergoing similar transitions to the subsequent levels of *learning*, characterized by the automatic adaptation of connections, *thinking*, and possibly even *metarationality*. Such transitions would dramatically increase the network's power, intelligence and overall usefulness. They can be facilitated by taking the 'network as brain' metaphor more seriously, turning it into a model of what a future global network might look like, and thus helping us to better design and control that future. In reference to the super-organism metaphor for society this model will be called the 'super-brain,'" Francis Paul Heylighen, Johan Bollen, "The world-wide web as a super-brain: From metaphor to model," 1996 (retrieved from http://pespmc1.vub.ac.be/papers/WWWSuperBRAIN.html on March 11, 2016); also F. Heylighen, "Cognitive levels of evolution," in: Felix Geyer (Ed.), *The cybernetics of complex systems*. Salinas CA, Intersystems,1991.

714 See Wendell Wallach, "Robot minds and human ethics: The need for a comprehensive model of moral decision making," *Journal of Ethics and Information Technology* 2010, vol. 12.

and "indirect reciprocity" between "genetically homogeneous"[715] bots and, in the future, between bots, humans and nonhuman living beings? We can imagine serving robots coming to the rescue of people and pets after an earthquake being 'obliged' by an imperative of altruism and "hostage" (in Levinas' terms), but we also know that altruism and empathy must be accompanied by rationality to create an efficient moral strategy. Following the developmental scale of Allen and Wallach, we can imagine bots that personify virtues, such as compassion,[716] on the one hand, but on the other hand "hard" cognitive and metacognitive competences such as operating the categorical imperative test. We can even imagine that a highly developed artificial moral agent does understand complex topics such as a transcendental *causa noumenon* which is unrelated to phenomenal reality, but is instead attributed with an intelligible "necessity" resulting in the highest moral self-obligation, vocalized in the formula of the categorical imperative. Contemporary unorthodox Kantians[717] try to exceed the narrow Kantian range of autonomous subjects in order to involve animals: a naturalized idea of animal autonomy follows. Since we witness similar developments in research on artificial moral minds[718] one may conclude that there is no need for naturalizing the term *autonomy* in AI. On an advanced level, as moral-cognitive theorists and experimentalists show, there is a need for high–quality normative criteria and standards of good practices. But high–quality criteria and a high number of criteria is not the same. As we read in Rosalind Picard:

> "The greater the freedom of a machine, the more it will need moral standards. I do not think designers will easily be able to enforce 'The Three Laws' (...) A system that truly operates in a complex and unpredictable environment will need more than laws; it will essentially need values and principles, a moral compass for guidance (...) Who has moral authority over computers, robots, software agents, and other computational things? This authority currently lies in the hands of those who design and program the computers.

715 D. Floreano et al., "Evolution of altruistic robots".

716 See James Hughes, "Compassionate AI and selfless robots: A Buddhist approach," in: P. Lin et al., *Robot ethics*, pp. 69–84.

717 E.g. Christine Korsgaard, "Fellow creatures: Kantian ethics and our duties to animals," *Tanner Lectures on Human Values* 2004, vol. 25.

718 Bradley J. Strawser "Moral predators: The duty to employ uninhabited aerial vehicles," *Journal of Military Ethics* 2010, vol. 9, no. 4; see also Jeroen van den Hoven, Gert-Jan Lokhorst, "Engineering and the problem of moral overload," *Science and Engineering Ethics* 2011, vol. 18, no. 1, pp. 143–155.

> Or, perhaps, in the hands of the one who provides their salaries, or the shareholders of the company, and so forth. Ultimately, it is a question for society as a whole"[719].

Or – in the light of developmental dynamics observed in the autonomous AI field – it is a question of fair, just, formal, rational[720]and – consequently – universal principles already known to reasonable humans. "Formal" means that such principles neither dictate nor prohibit concrete forms of behavior. Their role is very different, for they enable agents to examine the moral quality of a potential behavior before it is taken, and in accordance with all the potential autonomy personified in moral agents (whose list begins with human beings). On this point I disagree with Allen and Wallach, for whom formal principles such as Kant´s categorical imperative are too complex and unfeasible for artificial agents. Or, more exactly, their critics refer to the artificial agents labelled as embodying solely "operational" and "functional" moral competences. Let us examine how the categorical imperative test could work in the form of a simple algorithm.

4. A Categorical Imperative Test for Artificial Moral Agents?

As Immanuel Kant´s procedure of the categorical imperative is said to be not only rational, but also abstracted from all other practical moral principles, one may imagine the application of the categorical imperative by an artificial, intelligent, autopoietic system. "As engineers we implement these intellectual vehicles *back* into the world, for example as robots (. . .) Therefore we have a mutual interplay between the cognitive apparatus and the information it retrieves. Note that 'information' only makes sense for the individual who integrates the existing network

719 R. W. Picard, *Affective computing*, p. 134.

720 An AI "reasoning is based on rules, as opposed to the mixture of rules and feelings used by people", Picard continues. "It cannot *feel* what is most important to evaluate. The computer can explore more potentially meaningful relationships than a human, but it cannot yet feel which of all the possibilities are the most meaningful. Meaning is not obtained merely in associative connections; it is also accompanied by a literal feeling of significance", R.W. Picard, *Affective computing,* op. cit., p. 135. That is a good point since in my opinion, advances in sensitive AI design are too much concentrated on reading and imitating emotional states of living beings (social component), but they only scarcely focus on the epistemological role of moral emotions in moral reasoning and decision-making as a cognitive process (not only "personalized/impersonalized", vide Joshua Greene, Jonathan Haidt, "How (and where) does moral judgment work?," *TRENDS in Cognitive Sciences* 2002, vol. 6, no. 12.

of schemata"[721], that is, who is not only able to detect, gain and learn information from its environment (be it a social environment) and to process them, but also to create novel cognition which involves processed information and to operate (apply, respectively) it, as all intelligent systems including organic ones do as systems interconnected with their environments. One may at least examine if the categorical imperative, hypothetically translated to an algorithm, would be useful for an artificial intelligent system in the same way that it is (or could be) useful to intelligent human beings[722]. Both kinds of cognition, natural and artificial, need but a meaningful information about the action whose moral legitimacy (or claim for validity in terms of discourse ethics) is to be proved by means of the categorical imperative procedure. This meaningful information may be a 'maxim' containing a descriptive information on a relevant, sociomoral context (sociomoral environment) related action. Let us conduct a corresponding thought experiment.

On the other hand, "a maxim is the subjective principle of the volition"[723]. What does it mean when our individual "maxim" has not only some descriptive content, but also "moral" content? Moral content cannot be derived from the descriptive content such as the related sociomoral context of decision and action. It can be only 'authorized' as moral due to the categorical imperative test. Could I will that my maxim become mine and, potentially, also a "universal law"[724] for all? Who are the "all" then? Why do so with individual maxims which express our way of acting, maybe some habit, maybe some efficient strategy, or a "private" law? Why not rely on our own prudence, or just follow statutory laws? There are no private laws and the entirety of freedom cannot be governed by statutory laws. A substantial area is left for individual or interindividual governance. There are individual maxims which may have just material content or normatively valid material content, and there is a formal principle – a law – to test maxims to see whether they deserve such validity, or not. "I ask myself only: Can you also will that your maxim should become a universal law?"[725]. The maxims

721 Tom Ziemke, Alexander Riegler, "When is a cognitive system embodied?," *Cognitive Systems Research* 2002, vol. 3, pp. 342–344.

722 That cognition seems not to produce the 'epistemic feelings adjusting their cognitive operations', see Santiago Arango-Muñoz, "Two levels of meta-cognition," *Philosophia* 2011, vol. 39, no. 1, pp. 71–82; see also Bruce Wilshire, *Fashionable nihilism. A critique of analytic philosophy.* Albany, State University of New York Press, 2002.

723 Immanuel Kant, *Groundwork for the Metaphysics of Morals*. Trans. A. Wood. New Haven, London, Yale University Press, 2002, p. 16.

724 I. Kant, *Groundwork for the Metaphysics of Morals*, p. 18.

725 I. Kant, *Groundwork for the Metaphysics of Morals*, p. 19.

which deserve validity as being potentially universal are those which I ought to follow as a moral subject and decision maker. The "*pure* respect for the practical law is what constitutes duty"[726] as well as my identity resp. *self* as an autonomous *ethical* lawgiver.

When asking myself as an ethicist, why people use the categorical imperative exceptionally, I must agree with Kant: in past ages the moral subject could not fit with all her maxims "into a possible universal legislation"[727], thus Kant equipped the subject with a unique, supreme and very formal moral principle enabling her to examine her maxims and see whether they could potentially become universal *ethical laws*. Kant's ethical vocabulary is a pendant to his legal-theoretical vocabulary. Today, in the era of pluralism and diversity, a subject can easily find plenty of ethical laws and standards. In democratic cultures legislation corresponds to human autonomy and promotes the belief 'what is not prohibited is permitted'. This normative framework brings a release: one is not left to his or her own devices with one's own questionable maxims.

But will AI ever have sufficient access to ethical criteria for all the kinds of its actions, including "all the occurrences that might eventuate,"[728] as Kant puts it? Probably not. Human beings are in a similar situation. When facing novel or extremely challenging moral issues we all need principles which are "universal" in a way that allows us to apply them to various practical and, simultaneously, sociomoral contexts. In Kant's terms, it is "maxims" that articulate the purpose of intended actions and practices.

Hilary Putnam approached morality as a computational system of reasoning that is only possible for individuals. Kant's categorical imperative was originally too developed for individual use. Having reservations about the moral personhood (or moral agency) of AI, one may go beyond that distinction and, according to Jennifer Hornsby, suppose the impersonal status of AI: "From the *personal* point of view, an action is a person's doing something for a reason, and her doing it is found intelligible when we know the reason that led her to it. From the *impersonal* point of view, an action would be a link in a causal chain that could be viewed without paying any attention to people, the links being understood by reference to the world's causal working"[729]. There is nothing 'deterministic' or

726 I. Kant, *Groundwork for the Metaphysics of Morals*. p. 19.
727 I. Kant, *Groundwork for the Metaphysics of Morals*, p. 19.
728 I. Kant, *Groundwork for the Metaphysics of Morals*, p. 19.
729 Jennifer Hornsby, "Agency and causal explanation," in: Alfred L. Mele (Ed.), *The philosophy of action.* Oxford NY, Oxford University Press, 1997, p. 283.

'mechanical' in impersonal reasoning by following the categorical imperative as the core criterion of a maxim's moral legitimacy, providing this maxim with an obligatory claim. I would suggest Kant's categorical imperative procedure shows adequate transparency and objectivity to be applied by all kinds of individual agents in order to promote their ethical self-lawgiving. I can imagine an artificial intelligent agent applying it at least in an experimental context. I can imagine even more: namely, that, similarily to human individuals, such an individual artificial agent could become responsible for the broader social consequences of its activities as it conducts imperative-based reasoning. According to Kant, this reasoning must involve myself and other agents as *subjects* instead of objects (or any abstract entities). In all kinds of actions intended by myself I shall respect all agents which personify the ability to govern themselves in a reasonable and autonomous way, which is a synonym for their intrinsic and inalienable dignity, current or potential. In other words, I shall treat all these agents as subjects, persons, and "ends" in themselves (autotelic ends) and not as tools who can help me to achieve other goals, regardless of their nature. Such a "systematic union" of moral "lawgivers" regarded as autotelic ends is ruled by a universal moral principle and universalisable ethical laws as well. It is the preoriginal foundation of Kant's idea of the "Kingdom of Ends" whose core principle, embodied in all morally minded agents, at least potentially, says:

"Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction."

There are several versions of the categorical imperative in Kant, some of them more formal and less complex than others. This, however, does not imply that cognitively less advanced agents would be able to apply a categorical imperative test in an automatically *tacit* way. At this point, I would disagree with Harold Stone's argument, according to which "for people to follow the rules of an algorithm, the rules must be formulated so that they can be followed in a robot-like manner, that is, without the need for thought"[730]. Nowadays, we are facing a novel AI generation, e.g., machines that have begun thinking, and – unfortunately – humans that have stopped thinking.

A further problem with AI´s ethical reasoning would be the matter of the "will" and the will itself. How can an artificial intelligent agent "will" a potentially universal state of affairs which is normative by its very nature? It can only

730 Harold S. Stone, *Introduction to computer organization and data structures*, New York, McGraw-Hill, 1972; also Giulio Tononi, "Integrated Information Theory of Consciousness: An updated account," *Archives Italiennes de Biologie* 2012, vol. 150.

"will" something linked to the chain of its goals and purposes. Its "will" cannot be as intelligible and pure, e.g., oriented towards a moral duty as was postulated in Kant's philosophy. Thus, an artificial "will" needs to be replaced by a more formal term, e.g., another kind of causation than duty-based 'incentives', moral emotions, or even epistemic feelings. Such causation would originate from principles (or otherwise defined reasons) governing one's decision-making process. This resembles Donald Davidson's nomological approach to agency and action: "our justification for accepting a singular causal statement is that we have reason to believe an appropriate causal law exists"[731], Davidson states. "There must be a covering law," "though we do not know what it is,"[732] he continues. With regard to AI, to which an intuitionist approach does not apply, much more plausibility concerning moral instances as *governing laws, rules of grammar, logic,* etc. is expected. In other words, defining ethical procedures for AI, one cannot appeal either to the metaphysical attitudes of the AI nor to its 'intuition' or any deep-psychology related realities.

The next issue to consider would be a material determination of the maxim, e.g., the maxim's content made of situational contexts observed and learned by AI on its own[733]. According to Brian Tomasik, both kinds of problem should be considered (and maybe resolved) in the following way:

> "The categorical imperative makes most sense to me when interpreted through the lens of decision theory. In particular, compare Kant's formulation of the categorical imperative with this summary of timeless decision theory: Choose the output to your cognitive algorithm whereby you can at the same time will that it should become the universal output of all instances of that cognitive algorithm. This clears up the fuzziness about exactly what maxim our action is supposed to be following, since the 'maxim' is whatever algorithm we're executing when making a given decision. In fact, there are many

731 Donald Davidson, *Essays on actions and events*, Oxford NY, Clarendon Press, 2001, p. 160.

732 D. Davidson, *Essays on actions and events*, p. 160.

733 This corresponds to, and simultaneously goes beyond the contemporary concept of algorithm: "AI algorithms are usually only programmed to provide an answer based on the data they've learned. That is, we can see their conclusions, but most of the time we don't know how they arrived at them. That limits our ability to improve AI when something goes wrong, as well as learn from them when they make a decision that wouldn't occur to us", Dave Gershgorn, "We don't understand how AI make most decisions, so now algorithms are explaining themselves," *Quartz* 2016, December 20 (retrieved from https://qz.com/865357/we-dont-understand-how-ai-make-most-decisions-so-now-algorithms-are-explaining-themselves/ on May 18, 2018).

> algorithms that go into a given choice, so presumably we should act as though we're determining all of them at once. I don't know exactly how to make this work, but now we can see that it's just a technical problem in the realm of decision theory"[734].

Among various versions of the categorical imperative[735] there is one formula which focuses on the absolute respect for autonomy in all moral lawgivers. It seems to be useful for constructing an experimental ethical algorithm for AI. Similarily to its human users, such an algorithm could assist autonomous AI in demanding practical contexts where it has to make ethical decisions, but, at the same time, there is a lack of a superior normative criterion, a decisive rule, a standarized procedure, etc. or – alternatively – heterogeneous, conflicting norms handicap decision-making. There are controversial and dilemmatic issues, as yet unresolved problems, and novel challenges belonging to the practical contexts with such normative deficits. To construct a suitable model, several stages of algorithms would be essential:

(0) circumstances with respect to the practical context related algorithms able to detect, select, and qualify data (information) necessary to construct a descriptive (material) purpose of practical maxims;
(1) algorithms selecting morally relevant information in respect of the practical context;
(2) algorithms processing information in order to construct a maxim in a correct way;
(3) algorithms checking whether there is not a legitimate superior legislation, the main ethical context-related law/norm, and procedure to testify the maxim, and selecting out maxims testified by existing laws/norms (conclusion: maxims M^1 and M^2 are left for the categorical imperative procedure);
(4) algorithms operating the categorical imperative formula, such as for example 'Maxim M^1 is thinkable and – i.e., at least epistemologically correct – to become a rule acceptable to all autonomous agents including 'me', situated in analogous practical sociomoral contexts (conclusion: M^1 shall be observed at all analogous times regardless of alternatives, in Kant's terms – "pathological incentives").

734 Brian Tomasik, "Interpreting the categorical imperative," 2015 (retrieved from http://briantomasik.com/interpreting-the-categorical-imperative/ on April 8, 2018).

735 The hypothetical imperative will not be considered here for it is combined with a consequentialist approach. Furthermore, the formula 'you shall do A to achieve B' would require an ethical (categorical imperative based) test of both elements separately; the aim as well as the tool.

(5) in particularly socially sensitive circumstances, the algorithms which detect all related autonomous subjects and define them in terms of autotelic "ends", including natural and artificial agents.
(6) algorithms responsible for consulting all related autonomous agents and asking for their acceptance, negotiating their participation or contribution when an intended action is cooperative in nature, or it must involve persons' "conscious consent" typical for medical contexts.

I do not insist on this simplified categorical imperative procedure to be the sole criterion for ethical decision-making in AI. I do not even insist that it should be prior to all other ethical and metaethical procedures of providing moral reasonings with some consistency and transparency to make morally relevant choices and decisions legitimate in a universal way, as was originally thought in Kant's ethics for autonomous human agents. Certainly, Kant´s conception and the simplified categorical imperative procedure are not equivalent in meaning, especially since here autonomy is disconnected from the metaphysical notion of "Humanity" as being absolutely valuable, i.e., "whose existence in itself had an absolute worth," and its implications limited to humanity (accordingly, in the thought experiment conducted here all the autonomous agents' existence remains absolutely valuable). I merely suggest that statistical, mathematical, analytical, utilitarian, consequentialist, altruistic, empathic, virtue, etc. -based decision procedures are as less efficient among human agents, let alone artificial ones.

5. "No One Really Knows How the Most Advanced Algorithms Do What They Do"[736]

Teaching machines how to apply the categorical imperative test may have important implications not only for numerous fields such as medical care, military, security, management and investment decision-making where people rely on artificial intelligence agents. As already mentioned above, controversial, dilemmatic and novel challenges belong to them. "As deep-learning algorithms begin to set our life insurance rates and predict when we'll die, many AI experts are calling for more accountability around why those algorithms make the decisions

736 Will Knight, "The dark secret in the heart of AI. No one really knows how the most advanced algorithms do what they do. That could be a problem," *MIT Technology Review* 2017, April 11 (retrieved from https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ on April 11, 2018).

they do. After all, if a self-driving car kills someone, we'd want to know what happened"[737].

Risky developments involving decisions being made differently than human agents would make them could be prevented. Even algorithm and software developers only rarely understand how autonomous AI really processes its resolutions and reaches its conclusions, as Knight[738] stresses. Even less do we know about its ethical decision-making, including the question of whether it is integrated, hierarchical, etc., e.g., whether causal or logical interactions among always the same elements of the system occur, or whether they rather occur among alternative elements; or selected elements build lower and higher subsystems, as – hypothetically – a system of 'maxims' and a system of 'imperatives'; or whether a system of elements "generates a cause-effect information" which could be considered in moral terms[739]. For it shows that with regard to this secret attitude, an artificial mind resembles a "black box". Logical and deontological schemes such as the categorical imperative would enable humans to better track AI decision-making process and "to interrogate an AI system" (cf.) about its explanations, argumentations, and justifications in favor of or against a preferred conduct, especially in the face of novel and demanding contexts. Consequently, the reciprocal comprehension would also facilitate communication between natural and artificial intelligence and advances in the AI learning process as well. In this paper I argue in favor of understanding the complexity (and in favor of complexity as well) rather than in favor of the simplification of the AI's complexity in order to make it more transparent for human minds, for it would necessarily imply decreasing the benefits for humanity. I agree with Weinberger's argument:

> "Human-constructed models aim at reducing the variables to a set small enough for our Intellects to understand. Machine learning models can construct models that work (...) but that cannot be reduced enough for humans to understand or to explain them. This understandably concerns us. We think of these systems as making decisions, and we want to make sure they make the right moral decisions by doing what we do with humans: we ask for explanations that present the moral principles that were applied and the facts that led to them being applied that way. 'Why did you steal the apple?' can be justified and explained by saying 'Because it had been stolen from me,' 'It was poisoned

737 Dave Gershgorn, "The case against understanding why AI makes decisions," *Quartz* 2018, January 31 (retrieved from https://qz.com/1192977/the-case-against-understanding-why-ai-makes-decisions/ on May 6, 2018).

738 W. Knight, "The dark secret in the heart of AI".

739 G. Tononi, "Integrated," p. 297.

> and I didn't want anyone else to eat it' or 'Because I was hungry and I didn't have enough money to pay for it.' These explanations work by disputing the primacy of the principle that it's wrong to steal. It's thus natural for us to think about what principles we want to give our AI-based machines, and to puzzle through how they might be applied in particular cases. If you'd like to engage in these thought experiments, spend some time at MoralMachine.mit.edu where you'll be asked to make the sort of decision familiar from the Trolley Problem"[740],

but, not yet the sort of decision that is similar to the categorical imperative test. Currently, ethical algorithms are being developed and verified, in particular those concerning abduction. The latter can be defined as a "reasoning where one chooses from available hypotheses those that best explain the observed evidence, in some preferred sense"[741]. In the light of categorical imperative–based reasonings, the available maxims could be considered to finally choose that which most closely corresponds to a "preferred sense" expressed with the imperative. Pereira and Saptawijaya consider "representing moral facets by abduction" and "a priori integrity constraints (...) as a mechanism to generate immediate responses in deontological judgment"[742] as possible in AI. However, abductive reasoning based on the preferences applied in moral dilemmas advances mixed, e.g., the utilitarian and deontological ethics of AI at best. In so doing, researchers do not respect the a priori original meaning of the reasoning. Instead, they emphasize "the consequences of the considered abductibles have first to be computed, and only then are they evaluated to prefer the solution affording the greater good"[743]. As far as the categorical imperative procedure is concerned, the preference as well as the good are a priori well-known: it is all within moral agents' autonomy which potential conduct expressed with a maxim is to be validated as conforming to all the agents' autonomous self-governance or not. Further preferences, interests, goods, rights, etc. remain controlled by other kinds of procedure. I do not insist on the unlimited suitability of the categorical imperative. Other postconventional principles, such as the principle of not harming others, the Radbruch Formula, the respect and reciprocal recognition principle, the rule

740 David Weinberger, "Optimization over explanation. Maximizing the benefits of machine learning without sacrificing its intelligence," *Berkman Klein Center for Internet Society at Harvard University* (retrieved from https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d on Februar 1, 2018).

741 Luis Moniz Pereira, Ari Saptawijaya, *Programming machine ethics*, Switzerland, Springer International Publishing, 2016, p. 35.

742 L. Moniz Pereira, A. Saptawijaya, *Programming machine ethics*, p. 35–38.

743 L. Moniz Pereira, A. Saptawijaya, *Programming machine ethics*, p. 38.

of constitution, etc., rather belong to the basic principles of fair and just conduct. Utilitarian procedures seem not to be a sufficient warranty of those qualities of conduct in both humans and AI. "In humans, the individual understanding that there exists a self in relation to others makes possible participation in moral community"[744]. Unlike in humans, such an advanced and interactive self-understanding cannot be expected in artificial agents as they are, and, will most probably remain "rather simple-minded agents"[745]. For this reason, sociomoral principles and meta-principles, such as for example the unconditional categorical imperative, would certainly minimize restrictions and the harm inflicted on human or animal beings by AI-made self-willed solutions and decisions. At the same time, the AI's autonomy could get some novel dimensions, very different from a "slavish conformity to ethical rules"[746], as explained in previous sections. It does not so much need sophisticated ethical theory produced over the millennia. It rather needs a trained ability to make decisions in manifold practical and social contexts, when service, social assistance, companionship, and other relations with humans come into play.

According to Kant, the categorical imperative was thought to be supportive for those agents who face sociomorally relevant choices in contexts lacking normative guidance or ruled by unjust institutions. It also was thought to disclose, track and self-examine normative reasonings carried out by autonomously minded moral agents. Most probably, it was also supposed to be the most rational stage in humanity´s moral development. The main practical principle provides an agent with her own, autonomous and universally applicable deontological tool. In Kant, the procedure to examine own moral reasonings, called *conscientia*, also relies on the same deontological tool.

Designing autonomous AI nowadays, human agents try hard to instill principled self-determination into artificial agents. But sharing human autonomy with human-like agents results in challenges. They sometimes resemble interhuman challenges, such as for example trust and reliance on others. In fact, humans design autonomous AI in their own image to stay in relations with them. Humans stay in relations with others not just because the latter are autonomous, but able to control their activities within relationships. This kind of self-control implies respect and minimizes the risk. Otherwise, humans would not be able to trust

744 L. Moniz Pereira, A. Saptawijaya, *Programming machine ethics*, p. 163.

745 L. Moniz Pereira, A. Saptawijaya, *Programming machine ethics*, p. 164.

746 Paula Boddington, *Towards a code of ethics for artificial intelligence*, Springer International Publishing, 2017, p. 55.

their fellow humanoids as they do so with their fellow humans. Contemporary ethics often resists "Kantian attempts" to make social interrelations "a matter mainly of justifying universal principles for ideal rational agents"[747] whose observable performance would visualize at least the conclusions of their cognitive operations. According to Couzens Hoy, it also objects "to Hegelian attempts to assimilate otherness and to see the other as the mirror of the self"[748]. Perception seems to be the key concept in contemporary cognitive sciences and, most probably, it is an immediate communication tool between human and non-human agents including animals and intelligent devices. The verb *observe the rule* seems to link perceptual ability with intelligible apperception and following the norms together. "The key PCT contribution (...) is that human or animal organism controls neither its own behavior, nor external environmental variables, but rather *its own perceptions* (...) This fundamentally contradicts the classical notion of linear causation of behavior by stimuli (...) mediated by intervening cognitive processes"[749]. Tracking our own perceptions and other kinds of experiential and cognitive processes was already paradigmatic for transcendental philosophy ("Experience is cognition through connected perceptions", as Kant puts it in his first *Critique*). Transcendentalism questioned the mind-independent universe, in particular the moral universe, and explored all necessary preconditions of our cognitive access to it instead. It also revealed a unique, formal principle issued by practical reason, observable for all intelligent agents. This principle provides our motives and intentions expressed in "maxims" with at least potential shareable validity. Those intentions are strong enough to empower us as being cohabitants and fellow human beings to exercise our freedom, and thus our free will, in the real world.

However, unlike in human beings, "one 'special property' some believe is not to be found in any computational technology yet developed is free will. Conscious understanding is another."[750] Free will according to Kant is oriented towards the *normative meaning* of ethical and legal rules successfully examined by the

747 David Couzens Hoy, *Critical resistance. From poststructuralism to post-critique.* Cambridge, Mass., London, The MIT Press, 2005, p. 164 (despite the author's misinterpretation of Hegel's practical philosophy).

748 D. Couzens Hoy, *Critical resistance*, p. 164.

749 Vladimir G. Ivancevic, Darryn J. Reid, Michael J. Pilling, *Mathematics of autonomy. Mathematical methods for cyber-physical-cognitive systems*, New Jersey, World Scientific, 2017, p. 128

750 C. Allen, W. Wallach, *Moral machines. Teaching robots right from wrong*, p. 59.

categorical imperative test, and conceptualised to provide its user with their independent moral self-governance by legible and legitime practical rules within all sociomoral environments, even highly complex ones. Observing such a rule may be regarded as following an incentive (*unmittelbarer Bestimmungsgrund*), following a cognitive representation, or providing one's real judgment or decision with a reason based on a corresponding rule (*Vorstellung des Gesetzes*). Still, an autonomous lawgiver and a real autonomous agent (decision maker and action performer) are not the same.

An intelligible self-obligation (*Achtung*) and the will freely subordinating its autonomously given rule cannot be expected in AI's cognitive patterns even when they mirror the human ones. What is 'intelligible' in Kant cannot be reduced to mentalism or cognitivism, despite Davidson's efforts to explain transcendental activities in terms of "mental and nomological"[751] ones. "Consider, for example, Kant's contention that will and autonomy are necessary for an entity to be a moral agent. The ability to function as an autonomous being, or the capacity to will, suggest faculties beyond pure reason. However, little is understood regarding the manner in which the Kantian will and autonomy are supported by and emerge from the capacity to reason and other cognitive mechanisms"[752].

To stress one more time: observing practical rules given in an autonomous way which is imaginable for human as well as for artificial minds does not occur in a causal or deterministic way. Although, here we have to take note of the important distinction between rational procedures on the one hand and cognitive mechanisms on the other, both levels are considered to be "autonomous" by Kant[753]. A cognitive and rational moral agent deliberately *decides* to *act* in accordance (or discordance) to a rule, or other explicit normative criterion. Moreover, a mental moral agent feels obliged to follow it due to the rule´s imperative nature. Such an irreducible, metaphysical, intelligible self-commitment cannot be expected either in a cognitive system, be it natural or artificial. Here observing practical rules and acting accordingly may occur spontaneously, automatically and inexplicably. Both kinds of agents seem to make principled moral judgments and decisions; both of them make them every time *de novo.* None of them represents an autopilot-, routine-, and robotic-like rule-following mode.

751 Donald Davidson, "Mental events," *Philosophy of Psychology* 1970, pp. 208–225.

752 W. Wallach, "Robot minds and human ethics," pp. 245–246.

753 And maybe also by Wittgenstein. Though their concepts of cognition and cognitive process are different: Kant's concept refers to an embodied, while Wittgenstein's to a disembodied cognitive 'subject'.

The latter seems to be the most pragmatic and provident, but, in fact, it does not match the requirements of *autonomy* as a key attitude of the moral lawgiving subject and the 'inter-subject' constructed by Kant and forming the core foundation of moral sociability and society which could involve non-human intelligent beings.

We do not find our alter egos in those beings nor do we share essential *intelligible* faculties and principles funding the very reciprocity between us and them. Nonetheless, like Wittgenstein who was also advocating for the detranscendentalisation of rule making and rule following, we no longer need such foundations and explanations, but instead, a "*training* – comparable with the training you would give an animal"[754] and, vice versa, you would take of an animal just to create a novel kind of sociability, cooperation and community with them.

754 Rush Rees, Preface to Ludwig Wittgenstein's *Preliminary studies for the "Philosophical Investigations" generally known as The Blue and Brown Books*, New York, Harper & Row Publishers, 1978 (1st ed. 1958).