SIMONA TIRIBELLI

ARTIFICIAL AGENCY AND MORAL AGENCY: CONCEPTUALIZING THE RELATIONSHIP AND ITS ETHICAL IMPLICATIONS ON MORAL IDENTITY FORMATION

ABSTRACT: *ARTIFICIAL AGENCY AND MORAL AGENCY: CONCEPTUALIZING THE RELATIONSHIP AND ITS ETHICAL IMPLICATIONS ON MORAL IDENTITY FORMATION*
The paper aims to conceptualize the relationship between artificial agency, that is, the kind of agency connoting artificial intelligence (AI) systems, especially machine learning (ML) and deep learning (DL) algorithms, and moral agency, that is, the agency characterizing individuals as moral agents (persons), by highlighting its main ethical implications for moral identity formation. To this aim, after having unpacked the different forms of agency into question and clarified their main features, an ethical inquiry is carried out in order to show how artificial agency, as currently designed, not only interacts with, but it might also endanger the genuine development and expression of moral agency, by undermining individuals' epistemic and moral autonomy. The paper concludes by showing the main consequences that might arise from this impact, in particular in relation to those processes of individual and collective moral self-representation that nurture and steer the genuine formation and the flourishing of moral identity.

1. *Introduction*

The increasing advances in the design and use of artificial intelligence (AI), and especially of machine learning (ML) and deep learning (DL) algorithms, have prompted a robust corpus of literature in the field of philosophy and applied ethics over the last decade, assessing their opportunities and risks for individuals and societies. On the one hand, AI-based systems such as ML- and DL-based technologies have shown their huge capacity to perform more and more of our tasks and decisions efficiently

(i.e., quickly and at low costs) in many social domains[1], from communication and advertising to education, justice, and healthcare, by enabling their widespread conceptualization as a new «growing resource of interactive, autonomous, and self-learning agency»[2], along with their increasing delegation of – previously just human – choices, actions, and activities[3]. On the other hand, ML and DL algorithms, and thus, the new forms of "artificial agency"[4] they constitute, have been also widely unveiled to be very often flawed and/or biased[5] and to produce controversial effects, ranging from fostering social discrimination and inequalities[6] up to facilitating individuals' manipulation[7] and self-deceit[8]. In this scenario, whilst great attention has been paid to understanding and critically assessing whether ML and DL algorithms may be entitled to moral status[9], their interaction *with* and/or their impact *on* our moral agency have not been sufficiently explored thus far[10]. Nevertheless, this

[1] B.D. Mittelstadt *et al.*, *The Ethics of Algorithms: Mapping the Debate*, in «Big Data & Society», 3 (2), 2016.
[2] L. Floridi, *What the Near Future of Artificial Intelligence Could Be*, in «Philosophy & Technology», 32, 2019, p. 2.
[3] B.D. Mittelstadt *et al.*, *The Ethics of Algorithms*, cit.
[4] The concept of "artificial agency" is widely in use in the debate in ethics of AI to mean the capacity of AI systems (e.g., ML- and DL-based systems) to perform (also human) tasks and decisions (algorithmic decision-making) in a way that does not necessarily entail their further attribution of moral agency.
[5] A. Tsamados *et al.*, *The ethics of algorithms: key problems and solutions*, in «AI & Society» 37, 2002, pp. 215-230.
[6] C. O'Neil, Weapons of Math Destruction, Penguin, London 2016; V. Eubanks, *Automating Inequality*, St Martin's Publishing, New York 2018; S. Noble, *Algorithms of Oppression*, NYU Press, New York 2018; R. Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code*, Polity, Medford 2019.
[7] See the European "Declaration on the Manipulative Capabilities of Algorithmic Processes" (EU Decl [13/02/2019]).
[8] S. Natale, *Deceitful Media. Artificial Intelligence and the Social Life after the Turing Test*, Oxford University Press, Oxford 2021.
[9] S.M. Liao, *The Moral Status and Rights of Artificial Intelligence*, in «Ethics of Artificial Intelligence», Oxford University Press, Oxford (UK), 2020; L. Floridi, J. Sanders, *On the Morality of Artificial Agents*, in «Minds and Machines», 14, 2004, pp. 349-379.
[10] The majority of the discussion in the field of ethics and AI concerning moral agency is devoted to understanding artificial agency as morally connoted. AIs' interaction with human moral agency is less investigated and almost exclusively framed in relation to phenomena of social manipulation, surveillance and persuasive technology, and technological paternalism. See B.

inquiry sounds today needed considering how these systems, by deploying subtle and fine-grained techniques, have shown a huge capacity to capture, steer, and influence our preferences and beliefs[11] up to deeply reshape our decision-making processes and social and political[12] (including moral[13]) choices, by threatening our control and autonomy over them[14].

This paper aims to contribute to filling this gap and conceptualizing the relationship between *artificial agency* and *moral agency*. By artificial agency is meant the capacity of ML and DL that rule the functioning of today's majority of digital information and communication technology (ICTs) to "act" or "behave" (according to technical vocabulary)[15] by making decisions and accomplishing tasks in order to achieve certain pre-set goals. By moral agency is meant our capacity to develop genuine values, reasons, commitments, goals and moral ground-projects and to reflectively endorse them as motives for our choices, actions, and behavior. Specifically, the goal of the paper is to unpack this relationship and show how artificial agency can reshape but also hinder the genuine development of our moral agency and moral identity, by affecting the processes through which a) we self-

---

Frischmann, E. Selinger, *Re-Engineering Humanity*, Cambridge University Press, Cambridge 2018.

[11] L. Floridi, *The Fourth Revolution. How the Infosphere is Reshaping Human Reality*, Oxford University Press, Oxford 2014, p. 58; L. Royakkers *et al.*, *Societal and Ethical Issues of Digitalization*, in «Ethics and Information Technology», 20, p. 2.

[12] Consider the Facebook emotion experiment (2004) or the case of Cambridge Analytica (2016).

[13] On the impact of algorithms-based systems on our moral choices (and, specifically, on our freedom of choice), see S. Tiribelli. *Predeterminazione algoritmica e libertà di scelta*, in L. Alici e F. Miano (a cura di), *Etica nel Futuro*, Orthotes, Napoli 2020.

[14] BD. Mittelstadt *et al.*, *The Ethics of Algorithms*, cit., p. 9; M. Hildebrand, *Smart technologies and the end(s) of law: Novel entanglements of law and technology*, Cheltenham: Edward Elgar.

[15] This paper does not assess whether AI systems can be considered as moral agents, as the scope of this inquiry is limited to their interaction with individuals as moral agents. Thus, the use of the term "agency" (or "behavior") in relation to AI is that adopted in the technical debate: it refers to AI capacity to operationalize instructions (as in the case of deterministic algorithms) and/or to develop new paths to achieve a certain goal (as in the case of indeterministic algorithms).

represent and act as morally-connoted agents and b) we self-represent and act as part or members of a common-good-driven collectivity (a society sharing common substantial values). To this aim, a first section is devoted to introducing the specific features characterizing artificial agency and human moral agency. A second section clarifies how they mutually relate and interact (i.e., modality of inter-action) and how artificial agency, as currently designed, might endanger our moral agency by undermining our autonomy at the epistemic level (*epistemic autonomy*), thus, weakening us as *knowers*, and at the very stringent moral level (*moral autonomy*), weakening us as *moral agents*. A final section sheds light on further implications of artificial agency's impact on moral agency, by focusing especially on the role of the former in reshaping and/or hampering individual and collective moral self-representation processes which in turn shape the genuine formation and expression of our moral identity.

2. *Artificial Agency and Moral Agency*

The understanding and conceptualization of the relation between artificial agency and (human) moral agency cannot prescind from first clarifying the features mainly connoting each form of agency into question. Indeed, even if with the well-known "fourth revolution"[16], leading scholars have argued the blurring of the distinction between artificial agents and human agents on the basis of an unprecedented common capacity to process (acquire and act over) information, and being processed as information, when it comes to considering moral agency, this common broad definition of agency might result too partial and turn out to be misleading. Beyond the pretense of exhaustiveness of the wide critical debate on moral agency developed in moral philosophy, there is a widespread Kantian-inspired agreement on the fact that moral agency requires certain basic features of rationality, and

[16] L. Floridi, *The Fourth Revolution*, cit.

specifically, a certain degree of *moral reasoning* (that is, the capacity to form and reason on moral concepts, ideas of good, as well as moral reasons, values, and beliefs), especially when it comes to consider deliberative processes on alternative options that we might choose as reasons for our agency. Moral reasoning indeed allows the genuine formation and the exercise of the reflective endorsement of what can drive and motivate our agency (*moral motives*), from values and beliefs to shared goals, joint commitments, and moral ground projects, which normatively denote a mere agency as a morally-connoted one. Put it differently: moral reasoning allows the formation of our specific *moral knowledge*, that is, the formation of values, moral reasons, and ideas of good, and thus, broadly, of the moral motives that normatively steer our choosing and agency by developing a specific *ought to*. Therefore, moral reasoning enables *moral responsiveness* (both as *accountability* and *answerability*), that is, to offer others reasons (i.e., account) for our actions and be responsive to reasons provided by others. It sounds noteworthy underlying that in order to form a *genuine* moral knowledge, we need to be exposed to different ways of thinking, acting, and behaving (different moral practices), insofar as moral heterogeneity is crucial to assess whether our values, reasons, beliefs (and so forth) we endorse as moral rules for our choosing and agency are the optimal ones, or instead there are reasons to revise them. According this standard view, the genuine development and reflective endorsement of values, reasons, beliefs (…and so forth) as moral motives for our choices express the exercise of our autonomy and enable the genuine expression and flourishing of our moral agency. To sum up: moral reasoning is a basic feature of moral agency and it requires a heterogeneous moral exposure to enable truly the exercise of our *epistemic autonomy*, that is, to form our moral knowledge in a way that is genuine, as what we will endorse as moral motives (normative moral rules) for our behavior. Once we endorse those

moral motives and choose and act (or choose to do not act) on the basis of them, we exercise our *moral autonomy*, that lies in our reflective endorsement, and thus give shape to and express our moral agency. While the standard view on moral agency stresses the import of rationality features and moral reasoning, as it has been pointed out by theories of bounded rationality and studies in cognitive moral phycology[17], our deliberative processes and choices are not only a matter of rationality, but are also influenced and driven by emotional facts[18]. Hence, our moral agency and moral identity can be the result of choices also motivated by morally-loaded emotions. The latter, in turn, play a bivalent role, potentially having both a negative impact on moral agency, namely, counteract moral reasoning and endorsement, or instead motivating and informing it[19], and this mainly depend on what emotion is at stake and how that is triggered, thus, depend on particular forms that emotions can take in particular individuals. If our moral agency can be steered by motivations that involve the interplay of rationality and emotions, the role of the latter should be acknowledge to understand what characterizes our moral agency, especially when – as we will see later – the latter is considered in a new scenario of inter-action with artificial agency. When it comes to understanding artificial agency, the debate in ethics of AI very often uses a similar vocabulary to that pertaining to moral agency; however, substantial differences in meaning need to be highlighted. As anticipated before, by artificial agency is meant the capacity of

[17] On theories of bounded rationality, see D. Kahneman (2011), *Thinking, Fast and Slow*, Farrar, Straus & Giroux. For a contribution at the intersection of cognitive sciences and philosophy of mind, see M. De Caro, M. Marraffa, *Debunking the Pyramidal Mind: A Plea for Synergy Between Reason and Emotion*, in «Journal of Comparative Neurology», 8, 524, 2016.

[18] See A.E. Galeotti, *Political Self-Deception*, Cambridge University Press, Cambridge 2018; A. Mele, *Self-Deception Unmasked*, Princeton University Press 2000.

[19] On the role of positive and negative emotions for moral reasoning, see the contributions of M. Nussbaum, *Upheavels of Thought: The Intelligence of Emotions*, Cambridge University Press, 2003; *The Monarchy of Fear: A Philosopher Looks at Our Political Crisis*, Simon & Schuster, 2018.

ML and DL algorithms to make (also human) tasks and decisions to achieve certain preset goals. The latter are usually humanly decided, that is, pre-determined by designers according to technology providers' vision and, as widely documented[20], in line with third-party interests. To date, these goals, especially in ICTs, have been mainly driven to maximize utility and economic interest (e.g., revenue from ads); therefore, they have little to do with human moral goals or ground-projects. However, the recent rise of AI ethics as a discipline and movement is working to counteract this "hetero-determination"[21], by encouraging AI providers and designers to embrace a value-oriented approach to design AI systems in a way that respect ethical principles and foster the social good. Nonetheless, this virtuous scenario still requires some time to be soundly realized, and current goals of algorithms ruling ICTs are mainly preset to capture individuals' attention and maximize revenues resulting from clicks, purchases, and advertising and, therefore, they do not usually align with specific human (moral) goals, as well as no form of "autonomy" of these systems is exercised for the definition of such goals. In this regard, it is important to clarify that when we refer to artificial agency we mainly refer to probabilistic ML and DL algorithms, rather than deterministic (statistical) models, which capacity to perform tasks consists of a mere operationalization (automation) of instructions or rules humanly defined. Probabilistic or indeterministic algorithms show instead a heuristic and predictive power (i.e., *knowledge discovery method*) that expresses itself in the capacity to find, given a certain goal, existing or new paths to achieve it, by inferring predictive patterns or correlations. To sum up: they are not trained to achieve a certain goal but, given that, they can behave (deploy techniques) in order to *learn how* to achieve it. In this learning

[20] B. Frischmann, E. Selinger, *Re-Engineering Humanity*, cit.
[21] S. Tiribelli, *Predeterminazione algoritmica e libertà di scelta*, cit.

or "smart" feature lies what many scholars describe, with much criticism, as the algorithmic capacity of "reasoning" (or the "intelligent" side of these systems). Also, it is specifically in the consideration of this algorithmic capacity to self-learn (especially in unsupervised models) how to achieve certain goals that some scholars attribute a certain kind of "autonomy" to these systems – here too with not a few criticisms. Indeed, the more the algorithms prove the ability to find patterns that help to accurately predict how things will go even without instructions to follow, the more they are perceived as smart, autonomous, and thus powerful artificial agents. If the different features of the two kinds of agency into inquiry are now clearer, in the next section, we explore how artificial agents relate and interact with us as moral agents, by showing how the former, as currently behave, allow to conceptualize this specific new relation, more than as an inter-action, as a sort of over-action: how artificial agents more than *inter-act with* us seem to *over-act on* us as moral agents.

3. *AI over Epistemic and Moral Autonomy: who decides what?*

In the previous section, we have underlined how today we – as agents – are increasingly sharing our space – whose informational understanding (informational space: everything can be understood as information)[22] is now widely acknowledged – with artificial agents, with whom we share the capacity to unceasingly process and be fed by information, while, as previously clarified, we perform different kinds of agency. In this informational space, artificial agents, that is, ML and DL algorithms, sit in a privileged position both in knowledge and action in respect to ours both as knowers and agents. Firstly, they have the capacity to elaborate huge amounts of data, and also thanks to our continuous interaction with them (*unconscious human training*), semanticize them, by transforming vast streams of disaggregated and

[22] L. Floridi, *The Fourth Revolution*, cit.

heterogeneous data in valuable information, about us and the world, in the form of predictive patterns or correlations. In this sense, they have an *epistemic advantage*: they see what is invisible to us (including information about us) without being seen by us. This advantage is also a *practical* one, that is, it is a power asymmetry in knowledge and action: insofar they are informational gatekeepers, they can decide what option to show to us in informational terms; in other words: what *can* or *cannot* inform and motivate our agency. In this regard, that profiling and persuasive techniques are the means through which algorithms are mainly designed to achieve preset third-party (economic) goals is something today well documented[23]. Less explored instead is how they *inter*-act or *over*-act with/on our moral agency while they work to achieve their goals. Let us see it.

In mature informational societies, where our daily experience is almost ubiquitously mediated by digital ICTs and by the algorithms ruling them, the latter more than mediate are deeply reshaping our choice-contexts, the context where we form and make our choices and actions, that is, where we develop and express our moral agency. To clarify this claim, it suffices to think about how algorithms through filtering and classifying techniques determine which information (or informational options) we can get access to. These informational options do not only include trivial contents. Indeed, since algorithms today mediate and therefore embeds almost each aspect of our life (datification process), from affiliations, professional relationships, and political and religious orientation to values, reasons, and beliefs, informational options – algorithmically chosen – can embed almost everything that might become a motive for our action, including what we have defined as moral knowledge. In this sense, it sounds reasonable claiming that ML and DL algorithms increasingly decide what (also morally-connoted) informational option we can get access to and process –

[23] BD. Mittelstadt *et al.*, *The Ethics of Algorithms*, cit.

and adopt as a moral motive – for our choices and actions. More the algorithmic presence increases in our lives, more their agency on what can inform and drive our choosing and agency can become invasive. It follows that whether we experience the above described algorithmic action, we as moral agents do not only just interact with artificial agents, by providing them information on how understanding us and the world that they in turn process and re-use to feed us and influence our agency. This algorithmic agency that is designed toward the achievement of specific preset goals expresses itself as a conditioning third-party oriented pre-determination of our choice-contexts, limiting by pre-selecting what can or cannot become part of our moral knowledge (values, reasons, ideas of good, etc.). If we experience this pre-determining artificial action on our choice-contexts and moral knowledge we can fairly claim that our epistemic autonomy is endangered[24]. Indeed, to be threatened is our capacity to genuinely form and adopt – via critically assessment of – our moral motives (that we choose as normative rules for our agency) as the optimal ones. Indeed, algorithms tend to predetermine our informational context and exposure on the basis of subtle profiling and data mining techniques which process huge streams of data in order to create profiles as labels to attach to us, exploited in turn to categorize us in specific groups and to test on us (*fine-tuning*), as part of these groups, what preselected informational options (and contexts) work better in order to achieve their preset goals (from capturing human attention to increasing click and advertising or purchasing revenues). As documented in the debate[25], these ML techniques tend to foster filter bubbles or echo-chambers of profiled like-minded people, where phenomena like social polarization, social cascades, as well

[24] S. Tiribelli, *Predeterminazione algoritmica e libertà di scelta*, cit.
[25] C. Sunstein, *Democracy and the Internet*, In J. van den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy*, Cambridge University Press, pp. 93-110; E. Pariser, *The Filter Bubble*, Penguin, 2011.

as thought radicalization are fostered due to a lack of diverse points of view, including morally heterogeneous reasons, values, beliefs (and so forth). In this sense, to achieve third-party preset goals, algorithms have a considerable impact on our moral agency by over-acting on our capacity to form *genuine moral motives* that we can critically assess and endorse for our agency. However, our epistemic autonomy is not the sole to be endangered. Indeed, one might say that, even if our epistemic autonomy is affected, we can exercise our autonomy via the exercise of the reflective endorsement on available options, even if this means exercising our moral agency by embracing motives that are not developed in a truly genuine way (thus, critically assessed via a morally heterogeneous exposure). However, we claim that also our reflective endorsement, thus, our moral autonomy, might be algorithmically undermined.

In the previous section, we highlighted that emotions play an important role in motivating our agency and that their enabling and disabling role in the light of moral agency depends on the particular forms that emotions can take in particular individuals. Indeed, we as particular individuals are more than mere instances of rational agency (abstract selves): we are deeply connoted by specific attitudes, matured via affiliations, social relations, and experiences, as well as shaped by vulnerabilities, fears, and physical or psychological weaknesses. These traits connote us intimately, by shaping the way in which we live our life and rationally or emotionally respond to life events and other people. In other words, these personal traits shape how an event or information is perceived or interpreted and which kind of emotion might be generated in response. In this regard, that fine-grained algorithmic profiling techniques are able to capture these traits have been extensively shown.[26] However, algorithms, as currently designed, in order to achieve certain preset goal, do not only

[26] EU Decl [13/02/2019].

infer these traits and use them to sub-categorize us in groups to which show through filtering and classifying techniques preselected informational options rather than others, predetermining the epistemic conditions of our choosing and agency. Even more impactful techniques such as micro-targeting algorithmic recommender systems (RSs) are increasingly deployed to meet third-party goals. These work by exploiting inferred intimate traits (such as weakness or vulnerabilities) in order to trigger specific disruptive emotions, especially primary emotions (like fear, disgust, anger), which – even if functional to achieve preset goals – are more likely to disenable, rather than enabling, our moral reasoning and the exercise of our moral autonomy. In these cases, we might assist to a phenomenon we can define as the *suspension of the reflective endorsement*. Primary emotions indeed are defined in moral psychology as those we have in common with children, as they are instinctive and innate, and differ from secondary emotions that instead require a certain degree of awareness and socialization, along with the formation of an idea of good. RSs, by exploiting specific informational contents to target these specific individual traits, can elicit primary emotions, which in turn can trigger instinctive behavior-responses. The latter can suspend the exercise of our reflective endorsement, thus leading an algorithmically recommended emotionally-loaded informational option to determine our choices and actions at our own place. According to this standpoint, we claim that artificial agency, as currently designed, might undermine also our moral autonomy, specifically by deploying vulnerabilities-exploiting and primary-emotions-triggering techniques that can convert an – algorithmically recommended – informational option from being a *motive* (that inform and we can embrace) for our choices and actions to be the main *cause* of them, triggering instinctive or emotional choice-behavior responses that suspend our reflective endorsement. In this sense, we might not

only result as weakened as knowers, that is, in our capacity to reason and develop genuine moral motives (values, reasons, beliefs, etc.) to endorse in our agency, but also as moral agents, that is, in our deep normative capacity to endorse reflectively what can become a moral motive for our agency. It follows that our capacity to develop, express, and exercise our moral agency, more than just algorithmically informed, shaped, or distorted, might result deeply compromised.

4. *Conclusive Reflections: AI over Moral Identity Formation*

In the previous section, we have unpacked the relation between artificial agency, as currently resulting from the dense network of multiple algorithmic techniques that govern our pervasive ICTs, and our moral agency, showing how the former can undermine our epistemic and moral autonomy, and thus, more than *inter*-ac, can *over*-act on or *counter*-act the genuine expression and exercise of our moral agency. This final section sketches a few main implications of this impact for the formation of moral identity. Indeed, that digital ICTs reshape how we self-understand and constitute as individuals, hence, how we form our personal identity, has been already discussed in the debate[27]; less covered instead are the specific implications of algorithms for our moral identity's formation. We aim to shed light on a few of them in order to pave the way for further research on the topic. That artificial agency is currently shaping the social practices through which we develop our values, reasons, beliefs up to shared commitments and moral ground-projects should be now clearer. Indeed, it has been shown how algorithms affect our informational (and relational) exposure and pre-determine our choice-contexts in a way that tend to categorize individuals in *moral echo-chambers* where moral heterogeneity is diminished and people are incline to radicalize their beliefs and thoughts in order to be accepted by

[27] L. Floridi, *The Fourth Revolution*, cit.

other members of their group[28]. Also, we have argued how less social and moral heterogeneity (in thoughts, values, and reasons) tend to disenable moral reasoning and hamper the possibility to develop genuine moral motives which connote our agency as specific moral persons, that is, our moral agency. It follows reasonable claiming that artificial agency, as currently operating, is affecting our possibility to form over time genuine moral identities. Indeed, when we unconsciously end up in less heterogeneous and thus increasingly enclosed groups, formed through the exploitation of primary emotions (especially negative emotions), the capacity to critically assess if the reasons, beliefs, values (moral motives) we embrace are the optimal ones is eroded, due to the lack of a critical and heterogeneously informed dialogue, favoring a more emotional adherence to the claims made by members belonging to the same group (confirmation bias)[29]. This lack of heterogeneity and moral reasoning in turn undermine also our capacity to offer to others reasons (i.e., moral responsiveness) for our choosing and agency. It follows that algorithms not only reshape but might also hamper our processes of moral self-representation both at the individual and at the collective level: that is, how we self-represent and over time constitute as genuine moral agents (moral identity formation) both as individual persons and as member of a social collectivity that shares common substantial values and that is oriented toward the common good. It is likely indeed that the more our capacity to form genuine reasons and be responsive to different reasons is eroded, the more it is difficult to self-understand and self-represent as moral agents, by questioning if we are acting according to moral reasons, values, and projects that are aligned with the moral persons we would like to become. Conversely, we might find out – over time – that we have emotionally followed or

[28] Cass R. Sunstein, *Democracy and the Internet*, cit.
[29] C. O'Neil, *Weapons of Math Destruction*, cit.

approved practices that are disaligned to the kind of moral progress we would like to see in our society, experiencing a *moral disorientation* between who we are and who we wanted to be from a moral standpoint. This phenomenon can have also broader or collective implications. In fact, in like-minded and increasingly self-enclosed groups, along with thoughts' and beliefs' radicalization and the gradual erosion of the capacity to account for our actions, we might also become less responsive to others' diverse reasons, once we will encounter them, that is, less open and capable to understand and embrace other ways of thinking and acting (i.e., different moral practices). This phenomenon might lead us, as members belonging to different groups, to be less open and capable to achieve a potential agreement on topics of societal matter, thus, it might hinder our capacity to form collective joint commitments and share common values and goals – that is, to self-represent and act as part of a common-good-driven collectivity – which in turn might result in a diminished capacity of joint action and planning, which instead are crucial in order to create the conditions for a stable and morally flourishing society. It sounds the present as the right time to tackle or prevent these phenomena: revising algorithms' design in a way that considers the above-mentioned implications and allows the [re]alignment between moral agency and artificial agency is only the first step of a novel *ethics of inter-action* between *moral agents* and *artificial agents* that has just begun to take its first steps and imperatively demands to be further developed.

Simona Tiribelli è Ricercatrice in Filosofia Morale presso l'Università di Macerata e Research Affiliate presso il MIT Media Lab del Massachusetts Institute of Technology (USA)

simona.tiribelli@unimc.it