AI & SOCIETY (2023) 38:2045–2059
https://doi.org/10.1007/s00146-021-01313-x

OPEN FORUM

Interdependence as the key for an ethical artificial autonomy

Filippo Pianca[1] · **Vieri Giuliano Santucci**[2]

Received: 4 August 2021 / Accepted: 6 November 2021 / Published online: 10 January 2022

Abstract
Currently, the autonomy of artificial systems, robotic systems in particular, is certainly one of the most debated issues, both from the perspective of technological development and its social impact and ethical repercussions. While theoretical considerations often focus on scenarios far beyond what can be concretely hypothesized from the current state of the art, the term autonomy is still used in a vague or too general way. This reduces the possibilities of a punctual analysis of such an important issue, thus leading to often polarized positions (naive optimism or unfounded defeatism). The intent of this paper is to clarify what is meant by artificial autonomy, and what are the prerequisites that can allow the attribution of this characteristic to a robotic system. Starting from some concrete examples, we will try to indicate a way towards artificial autonomy that can hold together the advantages of developing adaptive and versatile systems with the management of the inevitable problems that this technology poses both from the viewpoint of safety and ethics. Our proposal is that a real artificial autonomy, especially if expressed in the social context, can only be achieved through interdependence with other social actors (human and otherwise), through continuous exchanges and interactions which, while allowing robots to explore the environment, guarantee the emergence of shared practices, behaviors, and ethical principles, which otherwise could not be imposed with a top-down approach, if not at the price of giving up the same artificial autonomy.

Keywords Artificial autonomy · Social robotics · Ethics · Interdependence · Intrinsic motivations

1 Introduction

Nowadays, one of the adjectives most frequently associated with robots and artificial intelligent systems is 'autonomous': it has become increasingly common to read and hear about autonomous artificial intelligence (AI), autonomous systems, autonomous machines, autonomous drones, autonomous weapons, autonomous robots, etc. News reports abound. Here is one of the many, almost daily examples: "A fully autonomous robot surgeon is the Holy Grail and many years off, says Dr. Tee, assistant professor of materials science and engineering at the National University of Singapore. He and other researchers are developing devices that can perform surgical tasks with minimal human oversight".[1] Although published in authoritative newspapers, the news is often hasty and generalist. This should not come as a surprise, given its target. One would expect to find more accurate clarifications (as well as a properly employed terminology) in articles issued by technical journals. However, at times, even here, some uncertainty is evident.

In many scientific publications dealing with theoretical aspects of AI and robotics, the terms 'autonomy' and 'autonomous' have recently become unavoidable. Concurrently, it has become common for these to be unspecified, used ambiguously, or only summarily defined.

While the discussion on artificial autonomy can be declined with respect to its natural counterpart, that is, how the autonomy of humans is preserved (or not) in situations of close interaction with machines (Anderson and Anderson 2010; Abràmoff et al. 2020; Bartneck et al. 2021), the core of our investigation is autonomy on the side of artificial agents: what does it mean for an AI or a robot to be autonomous?

✉ Filippo Pianca
filippopianca@icloud.com

✉ Vieri Giuliano Santucci
vieri.santucci@istc.cnr.it

[1] Advanced School in Artificial Intelligence, AI2Life, Institute of Cognitive Sciences and Technologies (ISTC), National Research Council (CNR), Via S. Martino della Battaglia 44, 00185 Rome, Italy

[2] Institute of Cognitive Sciences and Technologies (ISTC), National Research Council (CNR), Via S. Martino della Battaglia 44, 00185 Rome, Italy

[1] See https://www.wsj.com/articles/autonomous-robots-are-coming-to-the-operating-room-11599786000. Accessed 20 April 2021.

What does 'artificial autonomy' mean? Nowadays' robots can be considered 'artificial agents', i.e. machines that can reason inductively, make decisions, and act on the basis of a complex mechanism of sensors, algorithms, and actuators that makes them capable of physical consequences in the real world. For ease and convenience of explanation, we will refer in particular to contemporary social robotics: artificial agents equipped with AI and employed in the social realm will be the primary reference of this investigation, given the evident empirical repercussions of their actions.[2] The paradigmatic definition of robots both as "machines controlled by a software that move in physical space" (Tessier 2017: p. 179) and as "intelligent mechanical artefacts that can function autonomously" (Murphy 2019: p. 3) is widespread in the scientific literature. What we will try to clarify in this article is precisely the meaning of the adverb 'autonomously', which is mostly misunderstood and incorrectly defined, both from a technological and an ethical point of view.

The impact of social robotics—which is increasingly defined as 'autonomous'—on the world, on interpersonal relationships, and on society at large cannot but raise ethical questions. Indeed, each new technology raises issues on its development, distribution, and use, as well as on its ethical-anthropological implications. Autonomous robotics is no exception. Here, we will think of *ethos* as 'the place to live' (Reynolds 1993): it is up to us to decide where and how we want to dwell. Consistent with this reading, our focus will not be on distant prospects (whether utopian, rosy or catastrophic), but on the state of the art and the immediate future that current technologies seem to suggest.

The aim of this article is to propose a conceptual interpretation of the expression 'artificial autonomy', a theory on the meaning of robotic autonomy that can possibly support and validate some implementation suggestions already existing in the scientific literature while trying to address, from a new perspective, the social, cultural, and ethical issues deriving from this same expression. This purpose will be developed in three stages: (a) we want to clarify the meaning of the word 'autonomous' from the perspective of artificial agents, differentiating it from the concept of 'automatic' (Sect. 2); (b) after a brief mention of the hypotheses about superintelligence (Sect. 3), we intend to show concrete and current examples of artificial autonomy, evaluating both the reasons for its development and the near future perspectives that this concept presents us with (Sect. 4); (c) finally, we want to analyze what the ethical implications of this autonomy might be, trying to outline a strategy that manages to reconcile the advantages of adaptive and versatile systems with the need to ensure that artificial agents used in social contexts are safe and reliable (Sect. 5). All these issues are flanked by another, which is somewhat the key question of this work: do we really want autonomous artificial agents? Why? (Sect. 6).

2 Automation and autonomy

In the literature dealing with the impact of new technologies, and in particular that of artificial agents acting in the social sphere, the terms automation/automatic and autonomy/autonomous are often used interchangeably. Although it may seem superfluous to underline, these two semantic groups refer to different horizons of action (Chiodo 2021). A significant example of the muddled use of these concepts can be found in Galdon et al. (2021): although the authors first clarify unequivocally the difference between automation and autonomy, they then begin to refer to 'levels of automation' and 'levels of autonomy' interchangeably, thus revealing a semantic confusion. Nevertheless, the theoretic distinction between the two concepts as proposed by the authors is still accurate (even if later denied).

For the purposes of our reasoning, it will be useful to understand this difference. Let us focus briefly on the concept of automation. This can be defined by illustrating some of its characteristics: the presence of supervision (although perhaps very limited), the absence of learning or adaptation, the presence of a pre-programmed set of limited and precise tasks, the inability to choose new goals (intended as desired states or effects), and the low unpredictability (Galdon et al. 2021). In this sense, an automatic system is a pre-programmed device unable to learn from the context, to adapt to it (unless context-dependent instructions are explicitly coded into it), or to set its own goals; it is a system equipped with a fixed and finite set of specific tasks, whose actions are highly predictable (except for malfunctions). An automatic machine performs predetermined sequences of actions (Truszkowski et al. 2010)—just think of a driverless subway—and, therefore, demonstrates trivial behaviors (trivial from the perspective of autonomy, even if they can actually be extremely sophisticated). An automatic system is governed by prescriptive rules that allow no deviations (Tessier 2017), nor it has or can evolve its own motivations and purposes: it is not capable of intelligent problem-solving, given that the ability to learn is lacking (Castelfranchi and Falcone 2003).

From this set of negative characterizations, we can begin to deduce what autonomy means with respect to an artificial agent, i.e. to a robot. Let us try to introduce a first set of peculiarities of artificial autonomy, which will then be explored in Sect. 5. This preliminary portrayal derives

[2] By analogy, the discussion on artificial autonomy could also be adapted to virtual systems, that is, to all those sets of algorithms not implemented in a synthetic body: these too can have impacts on the real world, albeit mediated.

purely from the reversal of the characteristics of automatic systems. This idea is based on the assumption that automation and autonomy are two distinct and different, if not opposite, concepts (Chiodo 2021): while an automatic system is predetermined and unable to learn (it could be said that it is deterministic and schematic, in the sense of being deeply bound), an autonomous agent has no restrictive predeterminations (other than physical and normative ones) and can learn from the context in which it is placed.

Starting from the most relevant feature, an artificial agent capable of developing its own ends (final goals) should be considered fully autonomous. As we will see below, this aspect belongs to hypotheses about super-intelligences, that is, to a distant future (if ever knowledge and intention will allow us to get there). Staying at a lower level, closer to what is technologically possible, one could consider "beyond or out of automaticity" (Castelfranchi and Falcone 2003: p. 113) any artificial agent capable of performing reasoned actions on the basis of adaptation and learning. Furthermore, a robot can be considered autonomous if it does not need constant help from a human operator/user (it is not *continuously* supervised), if it adapts to complex environments and cooperates with other agents, and if it has an understanding of the world and a representation of the situation sufficiently elaborated to allow it to make decisions and act in a manner consistent with its ends (Tessier 2017). So far, these final goals have always been provided by the human operator or user. Overall, an autonomous artificial agent at the beginning of its career (that is, without having fully learned how to act properly) can be said to be quite unpredictable.

These characteristics, obtained preliminarily by contrast, reveal an issue that will be addressed in detail below, but which should already be mentioned: the question of the levels or degrees of artificial autonomy. Precisely in relation to the various limitations of an automatic system, we could say that the first level of artificial autonomy is that of learning, understood as the ability to solve certain tasks: if these are assigned, while the ways to deal with them are not openly explained, an artificial agent has to learn how to solve them. We can also think of robots that not only learn skills but can choose by themselves which tasks to focus on, starting from a predefined set. Finally, we can think of artificial agents who discover possible goals for themselves and choose which ones to dedicate themselves to.

Although all these characteristics are *analogically* consistent with the philosophical notion of autonomy—intended as 'self-rule', i.e. as what Kant calls "the property of the will by which it is a law to itself" (Kant 1997: vol. 4, pp. 440–441), namely the faculty by which one gives oneself a law—it seems that a founding element is still missing. A robot can be said to be autonomous if it is not strictly supervised (reduced monitoring is inevitable), if it is capable of learning and adapting, if it is not pre-programmed for highly specialized tasks (it can pursue quite general goals), and if it is rather unpredictable (this being an ambivalent feature). Is all this enough to explain the meaning of artificial autonomy? Addressing this question, most authors have relied on the notion of independence. Yet, far from corroborating the peculiarities introduced, this latter concept needs to be verified and integrated. Moreover, the issue of unpredictability raises another question: how can a robot operate in a socially adequate way, being autonomous?

3 Distant predictions: superintelligence and full autonomy

One of the key characteristics of robotic autonomy that has been introduced, if not the crucial one, is the ability of an artificial agent to develop its own final goals ('utility functions'). These are the ultimate ends towards which an agent carries out coherent and reasoned actions (intermediate ends). This capacity to decide by oneself one's ultimate goals can be defined as full autonomy (Totschnig 2020). With reference to intelligent artificial agents, the speculation is that in the future they may be able to autonomously choose their own final goals and act accordingly to achieve them (Bostrom 2014; Shanahan 2015; Lo Presti 2020). Some theses envision a general AI that will change and develop its own final purposes. To date, this ability is entirely hypothetical. By 'general AI' (the step before 'superintelligence'), we mean an algorithmic system capable of reproducing the complexity of human thought, that is, an intelligence capable of elaborating a total and semantic representation of both the world and itself. As it is well known, this is a hypothesis presented mainly to warn against advanced technological developments (Bostrom 2014). It will be useful to mention the relevant aspects of this debate, also by implicit comparison with humans.

The issue of artificial agents' full autonomy is an attractive but abstract logical exercise. While, according to some, intelligent artificial systems will never change their final goals, and therefore, will never exhibit full autonomy (e.g. Bostrom 2014), as this would be counterproductive and undesirable, according to others (e.g. Totschnig 2020) the finality argument (the fact that an artificial agent's final goals must remain such, i.e. definitive) is weak. Three objections are advanced by Totschnig: (1) as we humans happen to change our ultimate ends, reorienting our lives, even an artificial agent could theoretically be able to do so, if rational; (2) the ability to review one's final goals is in fact a sign of intelligence; (3) this re-examination depends on the development of the agent, i.e. on the advancement of its general understanding of the world (the improvement of its world representation influences its understanding of the final goals, and therefore, their transformation or adaptation). For this to

happen concretely, an artificial agent would need to have the cognitive tools suitable for learning from general experience (an aspect hitherto precluded). If a general AI or a superintelligence were able to learn and develop a complex worldview, no longer sectoral, by determining their own values and final goals, then they would demonstrate full autonomy (Totschnig 2020). As tempting as it is, this perspective is still far from what is expected on a technological level in the coming years: today's artificial agents have fixed utility functions established by programmers, just as their representation of the world is tied to specific domains of action. Any change of the final goals by the machine seems ruled out, for the moment.

4 Human independence: autonomous learning in artificial agents

The horizon of our research has to be reiterated: what artificial autonomy do today's technological skills and knowledge make possible? Given that prudential assumptions have an ethical relevance subordinated to their concrete realization, a gaze focused on what exists becomes a priority. By linking the discussion to the present and to the scientifically conceivable near future, we will introduce some concrete technological developments, as well as the impacts of their social implementation.

Full artificial autonomy, as the ability to decide or establish one's ultimate ends, is limited by several technical issues. Among these, artificial agents operate in specialized domains, i.e. they are applied in definite contexts, and consequently, it is not possible for them to construct complex and general representations of the world (Totschnig 2020). Additionally, artificial agents are currently bound to pursue the final goals set by programmers and developers. These purposes, however, are not necessarily specific. As we will see shortly, research on intrinsically motivated open-ended learning (Oudeyer and Kaplan 2009; Baldassarre and Mirolli 2013; Santucci et al. 2020) entrusts robots with rather general goals (e.g. 'be curious'). Although the programming aspect inevitably remains, it is minimized and no longer configured as a total predetermination in view of highly specialized tasks: thanks to a sophisticated collaboration between computational resources and sensors, intrinsic motivation signals allow robots to autonomously discover and set their own tasks while pursuing the quite general goal of increasing their knowledge and competences.

4.1 Machine learning approaches for autonomous artificial agents

Many artificial systems, both embodied and software-only, can be considered the direct descendants of Unimate, the first industrial robots produced by General Motors (Nof 1999): machines developed for well-identified tasks, operating in controlled environments, whose routines are (almost) fully programmed at design time (Groover 2016). Without the ability to learn new skills or adapt to the environment, these agents are limited to prescribed behaviors. However, a first example of 'independent' adaptation can be the one resulting from the interaction between the physical structure (and internal architectures) of the agents and the environment. From Braitenberg's vehicles (Braitenberg 1986), passing to Rodney Brooks' Nouvelle AI (Brooks 1991) to the recent Boston Dynamics automata, robotics has shown how 'simple' stimulus-action responses can generate a repertoire of extremely sophisticated and non-programmed actions: by letting the body structure of different simulated robots interact with the environment and its physical laws, it is possible to build "playful machines" (Der and Martius 2012) that, without learning, develop complex and unpredictable behaviors. Yet, if we want to look at artificial agents capable of acquiring increasing independence, we must focus on those strategies and algorithms that can guarantee both learning and reorganization of behaviors with respect to the context and the goals.

Machine learning is the branch of AI that deals with studying and developing algorithms that allow artificial agents to improve their performance through experience. Although we do not want to categorize in a clear-cut way the different types of learning strategies, and above all, we do not want to propose new scales or levels of autonomy that, as we will see later, are often arbitrary or too general, it might be useful to try to identify the different approaches according to their increased independence from the human programmer (or user). Supervised learning uses a pre-labeled dataset to train a system to solve categorization tasks. These techniques have been widely studied in the last years, thanks to significant advancements in the field of deep learning that brought unprecedented results such as those on image recognition (Krizhevsky et al. 2012). However, as their name clearly suggests, despite their power of generalization these algorithms necessitate a 'supervision' from the programmer both at the level of the assigned task and at the level of the solution (the labeling of the dataset) of that task. Differently, unsupervised learning deals with the discovery of hidden patterns or structures into collections of unlabelled datasets. This is the typical case of big data analysis, where AI systems draw connections and correlations between a huge amount of unstructured information, whose analysis would be unfeasible for human operators. Classic (but still widely used) examples are the so-called self-organizing maps (Kohonen 1990) and the K-means methods (Lloyd 1982): these are clustering algorithms that automatically organize unlabelled inputs on the basis of their features to build a description of the dataset. Although these techniques

present increased independence, they still rely on external interventions (which might also be from another artificial system) to obtain their dataset.

However, independence also lies in the ability to autonomously search and generate one's own inputs, so that a system's knowledge and responses grow cumulatively and in parallel with the development of the agent. This is the typical case of an artificial agent exploring the environment to acquire the necessary skills to solve a task. Learning by trial and error is the foundational feature of the reinforcement learning (RL) framework (Sutton and Barto 2018), one of the most popular (and bio-inspired) machine learning techniques. RL leverages the exploration of the states and actions spaces to allow an artificial agent to autonomously discover a policy (i.e. a sequence of actions) that given a certain environment maximizes the number of rewards provided for achieving a certain goal state. While the task (or, better, the reward function) is provided by the human programmer, the solution to the assigned problem is autonomously discovered by the system: in different words, we can say that in RL the final goal is set but the way to achieve it must be discovered by the artificial agent through the interaction with the environment, thus through the acquisition of new knowledge. Moreover, this knowledge might be context dependent, so that the system can learn different strategies to solve the same task in different environmental conditions.

RL and other machine learning strategies, allowing the autonomous discovery of solutions and behaviors potentially unknown at design time, guarantee further independence to artificial systems. However, these learning processes are always driven by specific assignments or requests (the tasks) that are coded by human programmers into the agents: while the 'how' can be human-independent, the 'what' is always heterodirected, even when some independence is allowed at the level of sub-goals (Barto and Mahadevan 2003). To use RL's terms, the reward function and thus the motivations of the systems are pre-programmed and strictly task-specific. Motivational autonomy, i.e. the ability to choose one's own goals, would be the real step toward artificial agents that are totally disengaged from human designers. Notwithstanding this scenario is currently almost unthinkable for nowadays robots, there is state-of-the-art research that in recent years has gradually raised the bar of artificial autonomy. Within machine learning and developmental robotics (Lungarella et al. 2003; Cangelosi and Schlesinger 2018), a field that tries "to model the development of increasingly complex cognitive processes in natural and artificial systems'' (Lungarella 2007), a new area called intrinsically motivated open-ended learning (Santucci et al. 2020) is producing a growing number of promising and cumulative results. The concept of intrinsic motivations (IMs) is borrowed from the literature on animals (White 1959) and from human psychology (Ryan and Deci 2000), describing how novel or unexpected 'neutral' stimuli, as well as the perception of control over the environment, determine learning processes even in the absence of assigned rewards or goals. In the computational literature, IMs have been implemented to foster different types of autonomous behaviors such as state-space exploration (Bellemare et al. 2016; Romero et al. 2020; Schillaci et al. 2020), knowledge gathering (Schmidhuber 2010), learning repertoire of skills (Singh et al. 2004; Oudeyer et al. 2013), affordance exploration and exploitation (Hart and Grupen 2013; Baldassarre et al. 2019; Manoury et al. 2019). Furthermore, and closely related to the topic discussed in this article, IMs have been used to allow embodied artificial agents to autonomously discover and select their own goals (Baranes and Oudeyer 2013; Santucci et al. 2016, 2019; Blaes et al. 2020). Instead of having specific tasks assigned to them, artificial agents are left free to explore the environment according to criteria such as novelty, unexpected events, or the improvement of their ability to achieve autonomously selected goals. Obviously, these criteria are implemented by human programmers, but asking robots to have as their ultimate goal to maximize a very general principle such as 'curiosity' makes these agents potentially unpredictable and free to develop and learn in unexpected directions. The aim of this line of research is to have versatile agents, able to interact with unknown environments and to accumulate knowledge and skills that can then be exploited to solve the tasks assigned to them. However, if we look at the studies on IMs from another point of view, we can identify in this research a first and concrete example of artificial autonomy, determined by the extreme generality of the ultimate goal with which robots are implemented.

Even if some research uses IMs in processes of imitation learning (Duminy et al. 2021), where robots are initially trained by human supervisors and then allowed to act autonomously, the majority of the research in this field is focused on robots and artificial systems that operate in environments where there is neither interaction nor the presence of humans.

4.2 Degrees of artificial independence

In principle and broadly speaking, if an artificial agent proves capable of expanding its learning, it can evolve a less limited understanding of the world and thus a sufficiently elaborated representation of the situation for it to plan, decide, and act in accordance with its overall goal. The actions it takes can be reasoned, that is, they can be based on adapting to contextual changes and on learning from experience (Redfield and Seto 2017). All this means that an autonomous artificial agent can demonstrate non-trivial behaviors (Grinbaum et al. 2017), i.e. actions that are not carried out schematically. Last but not least, an autonomous robot with a rather general goal, capable of a wider situational awareness,

of learning and adaptation, does not need to be *constantly* supervised by human operators (Defense Science Board 2012)—so long as it leads to more effective, comprehensive, and ethically adequate results than an automatic system. All these characteristics, achieved gradually, fit into a framework that can be called 'degrees of artificial autonomy'.

Taking up what has been introduced above, artificial agents that go 'beyond automation' seem to be characterized by different levels of artificial autonomy, depending on how far this distance from automation is extended. These levels are usually defined by adopting an increasing scale, precisely to highlight the gradual emancipation from automation. While it is true that "the overall effort to define levels of [artificial] autonomy has devolved into a philosophical argument" (Redfield and Seto 2017: p. 118)—meaning that researchers have tried to define frameworks that have usually turned out to be subjective and thus quite arbitrary, especially when put into practice—this attempt can still assist us in understanding how artificial autonomy is generally conceived. Once again, the meaning of what we seek will be achieved through opposition, or rather through integration.

The effort to define the levels of autonomy of sophisticated technological systems is not so recent. In a 1990 conference, Zeigler suggested a three-level hierarchy derived from model-based architectures. The first level involves the "ability to achieve prespecified objectives", so this would still be a system close to automation; the second level includes the "ability to adapt to major environment changes", and therefore, adaptation to contextual variations, as we have already seen, would be one of the first steps out of automaticity; the third level introduces the "ability to develop its own objectives", what we have called full artificial autonomy, a capability so far precluded to artificial agents (Zeigler 1990). This significant but narrow classification is based on a definition provided by NASA a few years earlier: "autonomy is the ability to function as an independent unit or element over an extended period of time […]" (Zeigler 1990: p. 4). Although NASA's purposes are extraterrestrial, and therefore not primarily social, they seem to suggest that a rising independence entails an increasing artificial autonomy. In short, the more independent a system is from human supervision for long periods of time, the more it performs set tasks adapting to environmental changes, the more autonomous it seems to be. Here is one of the first references to the correlation between the concepts of independence and artificial autonomy.

Some studies on autonomy levels for unmanned systems undertaken in the early 2000s have reiterated this connection. This effort has led to the development of ALFUS, a Framework For Autonomy Levels For Unmanned Systems (Huang et al. 2005a, b). In addition to proposing the creation of a common and shared lexicon, so as to facilitate the understanding of terms and the communication—a shareable suggestion that somehow informs the purpose of our article too—the framework provides two models: one detailed and one executive. The former establishes the parameters that determine the levels of autonomy of an unmanned system: among these, 'human independence'[3] (i.e. independence from the human) is fundamental (Huang et al. 2005a). The executive model—a linear scale from zero to ten—further confirms this correlation: the lower the interaction between human and machine, the more the latter can achieve a 'full and intelligent' autonomy for complex missions in extreme environments (represented by level 10). Analogous connections between autonomy and independence can be found in other taxonomies. For instance, Yang et al. (2017) claim that in fully autonomous medical robotics no human will be needed in the loop,[4] while Galdon et al. (2021: p. 206) state that a totally autonomous virtual assistant "can perform decisions solely on its own without reporting to the user".

Overall, being unstable and classifiable at will, no levels framework enjoys universal acceptance and sharing. Nonetheless, the widespread configuration of hierarchies provides a series of features that, regardless of their order, are constant. Among these, independence from the human stands out as the definitive parameter: the more functions are delegated to a machine, without supervision or under minimal control, the more this turns out to be autonomous (Murphy 2019). The goal of robotic design seems to be a progressive reduction of the dependence on humans, a continuous updating in the transition from a totally dependent robot to a fully independent one. In addition to causing ambiguity when put into practice—tasks can change dynamically—the idea of increasing levels of artificial autonomy seems to point out a willingness to equip robots with ever greater responsibilities. Indeed, in many texts dealing with this theme, artificial autonomy is also seen as the growing ability to understand and control a situation: the more a robot is *allowed* to take the initiative, the more it is considered autonomous (Murphy 2019). To summarize, in the literature that deals with the issue of levels of autonomy, the reasoning seems to be the following: progressive independence from the human realizes *tautologically* a growing artificial autonomy, which in turn implies ever greater levels of initiative and therefore more and more functions delegated to machines, which ultimately seem to acquire ever greater responsibilities. We could ask ourselves: is robotic autonomy really identifiable with the independence from the human?

The definitions of robot autonomy put forward regardless of the question of levels, on which there is no consensus, tend to underline this very same correlation. As Totschnig

[3] This parameter has become 'human interface' in the following paper (Huang et al. 2005b), a simply formal update.

[4] On this topic see also Yip and Das (2018).

(2020: p. 2473) recalls, "in the fields of artificial intelligence and robotics, the term 'autonomy' is generally used to mean the capacity of an artificial agent to operate independently of human guidance". This conception is also found in Tzafestas (2016: p. 196): "in robotics, autonomy is interpreted as independence of control". Others, such as Grinbaum et al. (2017: p. 141), argue that "robot autonomy is the capacity to operate independently from a human operator or from another machine". Some say that a robot that has the ability to work without external help, and therefore makes 'all its decisions' without human intervention, has to be considered autonomous and independent (Alaieri and Vellino 2016). In a nutshell, an artificial agent is commonly said to be autonomous if it is capable of choosing how to operate towards a goal (fixed but more or less broad) without the intervention and help of humans. From all these quotations, it emerges that the ability to be independent—conceptually reached first—determines the attribution of autonomy.

A first ambiguity is that both automatic systems and autonomous agents might perform processes independently, that is, without human intervention from start to finish (Truszkowski et al. 2010). This problem is quickly solved by making use of the difference between automation and autonomy discussed above: an automatic system executes predetermined and schematic sequences, and is therefore slightly controlled precisely because it is capable of operational independence in a limited context; on the contrary, an autonomous artificial agent is not guided by strong prescriptive rules and thus it can decide which tasks to focus on and which actions to undertake on the basis of learning and experience, with relatively little human intervention that allows it to adapt to wider contexts than the automatic colleague. Yet, this still does not explain artificial autonomy. Is it independence that realizes autonomy, or is it from autonomy that a gradual independence results? Redfield and Seto (2017: p. 104) argue that autonomy "enables the robot to operate with little or no human intervention to interpret sensor data or make decisions". If, through autonomy, robots 'are enabled to' act without much external help, that is independently, what is the basis of artificial autonomy itself? Being an ability, someone must first 'enable' robots so they can learn to be independent: how do artificial agents acquire the authority and means to act under less and less supervision?

5 Artificial autonomy as social interdependence

From what has been suggested, it seems that considering artificial autonomy only as the ability to operate 'as an independent unit or element over an extended period of time' is not enough. Although independence—understood here, from the viewpoint of robotics, as the ability to act with the aim of carrying out some predefined tasks without any external intervention or human help—can be one of the characteristics and consequences of artificial autonomy, it is not what appears to establish or explain this very same capacity. Even philosophical language, to which we have referred extensively in this paper, specifies a substantial distinction between autonomy and independence, as the latter postulates (almost in an individualistic or anarchic way) a rejection of bonds, affiliations, rules, and regulations, while this does not happen for autonomy (its etymology, after all, recalls the presence of a law, which is possibly moral and to be shared interpersonally). The difference lies in the relationship: while independence is the denial of any contact (except those strictly indispensable), and refers to a way of thinking, deciding, and living detached from any subordination to external authority and judgments, autonomy is philosophically based on relationships of mutual listening and interaction, and therefore calls for a moral coexistence which is the source of a common good.

If so, the definitions of robot autonomy based on the concept of independence are not satisfactory. Artificial autonomy requires, preventively, being able to learn expertise and skills (including social norms) from the environment and context. Such complex learning, a *sine qua non* of the capacity under consideration, seems possible only through the interaction with human agents and with the environment in which the robot operates. This is the reason why independence as such (understood as the denial of any dependence, law, or relationship) cannot ground robot autonomy.

In non-social contexts (e.g. exploration of the seabed or other planets such as Mars—see Washington et al. 1999; Huet and Mastroddi 2016), a direct relationship with other agents is limited or excluded, and therefore a robot would not cause problems (if not from an economic and operational point of view: it breaks down, it does not do what we would like, it wastes time, it destroys things that could have been used, etc.). Regardless of any extra-social purpose, be it aquatic or extra-terrestrial, and with renewed attention to social robotics (we are building more and more autonomous machines precisely to introduce them in different 'terrestrial' areas such as homes, hospitals, streets, schools, etc.), having artificial agents endowed with behavioral and motivational autonomy could lead to unexpected and potentially threatening prospects: damage to things and people, injuries, impairments, killings, but also moral dilemmas that question traditional human values.

Faced with these problems, a solution could be the abandonment of any development project. A naive, deleterious alternative: both because there are good reasons to build artificial agents that are not simply automatic but autonomous (think of versatility, the ability to operate effectively in unknown contexts, the capacity to find unexpected solutions,

more efficient cooperation), and because the constructive trend is already a well-advanced reality (Wallach et al. 2008) and as such it requires that problems raised by artificial autonomy are concretely addressed. But being unstoppable does not mean being unaware. How can we build artificial agents endowed with a certain degree of autonomy and at the same time capable of respecting social norms and ethical principles? Put it differently, how can an autonomous artificial agent behave ethically, that is, in a socially appropriate, safe, and respectful way?

These kinds of questions come from the 'machine ethics' perspective. As the Stanford Encyclopedia of Philosophy states, "machine ethics is ethics for machines, for 'ethical machines', for machines as subjects, rather than for the human use of machines as objects".[5] It is a research field interested in building intelligent agents capable of behaving in an ethically acceptable manner towards humans and other beings (Anderson and Anderson 2007). Broadly speaking, machine ethics (also known as artificial morality) deals with the design of robotic systems capable of demonstrating sensitivity to human societal and ethical values (Dignum 2018) and of adopting them in making decisions in morally significant contexts (Wallach and Asaro 2017).

An initial, feeble response to the aforementioned questions could propose limiting the range of action of social robots, placing them in totally definable, describable, and controlled situations. In this way, the proposal for artificial autonomy would become contradictory. Another solution could be a broad a priori implementation of ethical principles, a long (as desired) list of prescriptions to be translated into mathematical/logical formalism and to be applied in all possible scenarios (besides Asimov's infamous laws, see Gips 1995; Wallach 2008; Anderson and Anderson 2010;[6] Powers 2011). The code of an artificial agent would then include ex ante rules such as 'do not kill and do not harm sentient beings in any form' (these, clearly, should be much more specific and numerous). Regardless of the actual computational translation, it is unclear how such an attempt dropped from above could guarantee the social adequacy of the robot, making it a versatile moral machine.

Several doubts emerge: first, principles may be at odds with one another (Tzafestas 2016). There could be inconsistency between divergent indications and therefore operational paralysis, if not rambling actions. Furthermore, a purely deontological top-down approach could never identify all the ethical norms for robots, especially where these are context and culture-specific. Some rules are only tacitly expressed when respected (or transgressed) by humans: morality can never be fully specified.[7] If, in addition to moral rules, one tried to picture all the general prototypes of possible situations, this incompleteness would be even greater: the exhaustiveness of potential scenarios and behaviors cannot be predicted ex ante (Muehlhauser and Helm 2012). The main issue, however, concerns artificial autonomy itself. Providing a robot with many a priori stringent rules, and possibly pre-set situational patterns, would hinder any autonomous attempt. Indeed, in the top-down approach, the installed algorithms produce predictable results: programmers embed in the machine what they consider to be ethical behaviors, and the robot only has to determine when to apply them (Alaieri and Vellino 2016). As we have seen, high predictability is a characteristic of automatic systems: in adopting this approach with artificial agents, we would still face a schematic predetermination. If the goal is to build autonomous robots that are increasingly independent and adaptable, the path of high predictability has to be excluded. While it remains necessary to equip artificial agents with some fundamental ethical principle, the all-encompassing ex ante approach does not seem to be suitable: autonomy is not made up of preventive formulas. At this point, our question must be reiterated: how do we solve operational and ethical problems while guaranteeing the possibility of artificial autonomy?

5.1 Interdependence as a bottom-up approach to machine ethics

Despite robots cannot (yet) give themselves the law, i.e. decide their own final goals by themselves, and therefore be fully autonomous, they are somehow always in relationship with humans. In particular, this is true for social robotics (which is the primary reference of our analysis), whose purpose is to build intelligent machines to be used in different social contexts and at various levels of human interaction. If autonomy is a relational notion (Castelfranchi and Falcone 2003), and if this relationship occurs in a social context, artificial autonomy can only result from social interdependence. Upstream, in the definition of robot autonomy, there cannot already be independence. If that were the case, there would be no adequate contextual learning. When a robot is intended within a social environment, its autonomy must also be: an artificial agent, as an active entity, relates to a context in which there are other active entities, primarily human. Far from quickly becoming independent of them, a robot has to interact with humans in order to improve its

[5] See https://plato.stanford.edu/entries/ethics-ai/#MachEthi.

[6] Aldebaran Robotics' Nao—which in Anderson and Anderson's 2010 article is portrayed with a nimbus—is claimed to be the first robot implemented with an ethical principle.

[7] As Santos-Lang (2012) claims, "any rules we make [to be implemented in robots] will be imperfect, even if supplemented by moral intuition".

decisions and 'moral' expertise: its artificial autonomy is built interdependently. This means that, before being able to act socially without being supervised (or, more likely, under reduced control), that is, before being independent, an artificial agent must learn how to act, what is normatively adequate or inadequate to do in certain social contexts, even at intercultural levels. Such acquisition relies on interdependence: this concept seems to guarantee the possibility of artificial autonomy while safeguarding the ethical aspect. We understand interdependence as a mutual, social, and contextual dependence, a reciprocal dependence[8] so that robots can learn how to operate in an increasingly opportune way and at the same time humans can employ them for wider and more complex tasks, within increasingly less structured scenarios. One depends on the other for learning and autonomy, although obviously control is still on one side (however mild this can be).

Artificial autonomy is based on social relationships: the concept of interdependence reminds us that "a [social] robot is never isolated and that the human is always involved in some way" (Tessier 2017: p. 182). Relationships, however, are dynamic by definition: artificial autonomy belongs to a relational continuum (Defense Science Board 2012), i.e. it depends on the context of action, the task assigned, and the operational capabilities of the artificial agent. Put it differently, artificial autonomy is granted by humans to different degrees according to need: it is in the social relationship that robots are enabled to be more or less autonomous. It is the human, be it operator or user,[9] who allows the artificial agent to exercise a certain level of initiative. Similarly, though not entirely, Murphy (2019) speaks of an "adjustable autonomy" in which the human dynamically adjusts the autonomy levels of the robot.

Artificial autonomy represents a set of delegated capabilities (Defense Science Board 2012): depending on the situation, on its delicacy and vulnerability, as well as on the robotic design, a larger or smaller set of skills and tasks can be delegated to a machine with progressively less supervision. This shows that the autonomy we are looking for "is not an intrinsic property" (Tessier 2017: p. 182) of an artificial agent: as the robot is able to learn from experience and from a particular social context interacting with humans, its artificial autonomy is granted to always different (potentially ever wider) 'circles'. This is why artificial autonomy dwells in social collaborations based on interdependence. The same happens in the human dimension, which, after all, is the conceptual model for robotic autonomy: "genuine autonomy resides in the interaction between individuals and society. […] It is in this dialectical relation between the social and the individual that real human autonomy resides" (Dupré 2001: p. 18).

The biggest misconception that revolves around the concept of artificial autonomy is that robots can be completely independent of human control. Unfortunately, oversight is inescapable, albeit high-level (Defense Science Board 2012), as autonomous systems programming embodies the design limitations of decisions and actions delegated to machines, however, general and wide their goals may be. This means that the same artificial agent can be more autonomous in one situation than in another. This is why the issue of levels is both ambiguous and impractical: it is not a question of producing robots that are fully independent regardless of the context of operation, it is a question of building artificial agents capable of learning both from the specific environment in which they are located and from the entities they interact with. The point is to elaborate and develop the capacity for robotic interdependence: learning in collaboration, that is, maintaining interactivity during cooperation (Castelfranchi and Falcone 2003). This applies to delegation too, which does not mean totally uncontrolled reliance. Adopting the issue of autonomy levels as an operational reference implies that full autonomy, i.e. total independence, is always the most desirable end of robotic development (see Murphy 2019: p. 77): would not it be better to recognize that the degree of artificial autonomy has to dynamically depend on the task, the situation, the socio-cultural context, and the architecture of the robot?

The computational and material structure of the artificial agents that we want to be autonomous must be carefully considered: despite an undeniable sophistication, capable of overcoming human abilities under certain aspects, the implemented algorithms and the synthetic body remain operationally limited. The decisions and actions of a robot, however, general its purpose may be, cannot demonstrate absolute autonomy. Interdependence, as the inevitable relation existing between social actors, resizes robotic activities, establishing an independence of 'thought' and action that is always bounded (Defense Science Board 2012). This means that, in interdependence, artificial autonomy and consequent independence are never outright. Robots, far from being self-governing or autarchic, can only be "partially autonomous agents" (Tzafestas 2016: p. 2) since they have structural limits and cannot decide their own ends for themselves: their autonomy is defined only in relation to the

[8] The concept of reciprocity, in robotics, can be ambiguous (Van Wynsberghe 2021). In our paper it is intended conceptually: interdependence means reciprocal dependence, i.e. the robot depends on the human for its moral learning, while the human depends on the autonomous robot for its increasing adaptability. In addition to the fact that the human remains central, here reciprocity is instrumental (it is functional to have a versatile robot), therefore, it is not assumed in its full human significance.

[9] Clearly, a distinction should be made between operators and users on the basis of the technical knowledge possessed (Grinbaum et al. 2017).

decisions and actions taken to complete the required tasks.[10] Artificial autonomy, therefore, is an instrumental autonomy: a robot can decide how to behave in order to achieve pre-established human goals, be they specific or general. This agent, unlike an imaginary general AI, is autonomous but at the same time human-bound (Lo Presti 2020). In the social interdependence, a robot must be able to act autonomously but adequately by achieving objectives that are entrusted to it externally: its autonomy is executive, it concerns the choice of means, of intermediate and instrumental ends (sub-goals), and not of ultimate ends (see Castelfranchi and Falcone 2003: p. 106).

Artificial agents can be autonomous precisely, because they are related to (in a relationship with) humans: the mutual dependence that is established through the social implementation highlights that artificial autonomy cannot be absolute freedom. In analogy with humans, machines need constraints to be able to exercise autonomy. Beyond any logical paradox, the concepts of autonomy and dependence are compatible (Coeckelbergh 2006), as they are united by that of relationality: this is the dimension that establishes not only the advantages, but also the necessary limitations—which are ethical, not just structural, in the case of social robotics. Artificial agents can be autonomous while still in a social dependence that ethically delimits their action: not all types of external control threaten autonomy. The point for a robot is to learn these ethical limitations, respect them, and put them into practice. Indeed, the more autonomy is granted to a robot, the more "ethical sensitivity" is required (Tzafestas 2016: p. 66) not only from the developer and operator but also and above all from the robot itself.

Let us resume a question raised earlier: is it enough to implement ex ante moral rules in its code for a robot to be reliable and autonomous at the same time? As we have proposed, some ethical norms implemented a priori, i.e. during the design stage, are necessary (those that impose not to kill or harm sentient beings). An extensive top-down approach, however, would prove incoherent with the goal of building autonomous artificial agents, as we would fall back into automation. Simply put, ex ante basic ethical rules are necessary but insufficient. This being the case, it seems that these a priori norms need to be accompanied by a bottom-up approach in order to have ethical autonomous robots. It is, therefore, a question of considering a hybrid morality (Allen et al. 2005; Wallach et al. 2010), a sophisticated moral sensibility: the fundamental rules embedded a priori has to be supplemented by the learning of others, as well as by the learning of the ability to understand in which contexts to respect and exercise them. Thus, we would have robots endowed with an ethical autonomy that would allow them to be dynamic, flexible, and righteous.

The proposal that artificial autonomy should be based on social interdependence drives the development of this sophisticated moral sensibility too: put it differently, hybrid morality is the actual concretization—or, if preferred, realization—of artificial autonomy *theoretically* conceived as deriving from social interdependence. While, on one hand, the issue of the hybrid approach appears to be an inevitable and concrete consequence of the proposal advanced in this paper (here we quote it mainly to support what we suggest, seeking at the same time to understand how robotic interdependence could be implemented), on the other hand, the idea of interdependence as the source of robotic autonomy can justify and validate the hybrid approach itself.

Just as robotic autonomy (both behavioral and motivational) is achieved by interacting with the human environment, so the acquisition and operational understanding of broad ethical norms can only arise from the relationship with other moral agents, primarily humans. In a context of interdependence, artificial agents—through a sophisticated interweaving of sensors, deep and reinforcement learning algorithms, natural language processing, actuators, and so forth—can learn specific moral rules in a gradual and cumulative way by interacting with moral biological agents. This perspective is also relevant at an intercultural level (Dignum 2018), where the programming of a robot takes place in one socio-cultural context and its application in another. The concept of interdependence, therefore, can solve ethical problems while ensuring the possibility of artificial autonomy. This notion implies the introduction of a bottom-up approach in which "the programmer builds an open-ended system that is able to collect information from its environment, to predict the outcomes of its actions, to select among alternatives and, most importantly, has the capacity to learn from its experience" (Alaieri and Vellino 2016: p. 161). Robots endowed with this hybrid morality are able to learn from their attempts and errors, from experience, and from the surrounding environment (unsupervised learning) while keeping on relating to the pre-established fundamental principles (supervised learning).

Clearly, there is no lack of challenges: in this mixed moral perspective, robots should be able to adequately address the moral issues encountered in the interactions with humans, that is, they should interpret the moral relevance of situations and actions, formulate moral judgments, and communicate on morality (Malle and Scheutz 2014). Moreover, how to integrate different moral philosophies and dissimilar architectures? Probably, the ability to make moral decisions will require "some form of emotions, consciousness, a theory of mind, an understanding of the semantic content of symbols" (Allen et al. 2005: p. 154; Tessier 2017: p. 190),

[10] According to Tzafestas (2016: p. 2), "autonomy in machines and robots should be used in a narrower sense than humans (i.e., metaphorically)".

as well as the ability to grasp what is culturally meaningful and appropriate in each and every social context of operation. Finding an answer to these limitations, however, is not the purpose of this article—although framing artificial autonomy as interdependence could help in finding some solutions: a robot could learn how to manage variability and moral differences precisely through interaction and mutual dependence. For a more detailed explanation, we refer to future research.

While interdependence may place ethical limitations on the autonomous operation of artificial agents, it nevertheless envisages collaboration and cooperation capable of leading to more incisive and effective results than actions undertaken in full independence. A social robot, never being completely autonomous and independent, can behave more successfully if it is part of a group, if it helps it and at the same time learns better how to do it, being helped. Social actors, be they humans or artificial agents, make up a team: the more interdependent this group is, the more successful it will be. Indeed, the best teams are highly interdependent: together, robots and humans can achieve higher levels of innovation and better decisions, as well as reduce errors (Lawless et al. 2019; see also Lawless and Sofge 2017). If artificial autonomy results from interdependence, this dimension not only establishes moral constraints (Arkin et al. 2012), but also and above all social gains.

Mutual social dependence means that artificial autonomy can be seen as authority sharing.[11] Where human capabilities are limited for biological or cognitive reasons, an artificial agent can complement them, for instance by seeing more accurately or by operating in dangerous contexts. The benefits of artificial autonomy are evident (Redfield and Seto 2017). Nevertheless, as we have already mentioned, robotic abilities and decisions are limited too, as they are the result of algorithmic computations often modeled on a compromise between quality of the outcome and speed of calculation (Tessier 2017). These mutual limits can complement each other towards more fruitful actions: decision-making authority is shared, and so artificial autonomy integrates human fallibility, just as human cleverness compensates robotic constraints. With regard to the artificial agent, its autonomy has to be considered in a framework of authority sharing with the operator or user: the robot is allowed to take certain decisions and actions on the basis of its adequate learning, but never in a fully independent way. As for the human, it should not always be considered as "the last resort" (Tessier 2017: p. 186): human beings are prone to making mistakes or overestimating robotic decisions (overconfidence), as well as to intentionally hurting. Although authority sharing involves issues still to be solved (for example: which decision prevails? Who can act and when? In the event of a conflict between decisions, must the human always have the last word?), it nevertheless confirms the link between interdependence and artificial autonomy: where the human is limited, for whatever reason, granting an artificial agent a certain degree of decision-making and operational authority within social relations means making it autonomous, albeit never completely. Again, the fact remains that the robot has to learn how to decide and act ethically.

Interdependence guarantees the possibility of artificial autonomy: through contextual relationships a social robot can learn better and better how to act and behave appropriately. If it were primarily independent, without interactional capacity, its autonomy would be empty: it would rather be a different form of automation. The peculiarities of artificial autonomy, which we introduced above negatively, now find an explanation in the dependence existing at the social level between human and artificial agents. Postulating that a social robot is always equipped with machine learning algorithms in dialogue with sensors, actuators, and effectors, its ability to perform reasoned and non-trivial actions derives from its being in relationship with human agents: its computational structure and its synthetic-sensory body, constituting a sophisticated whole, ensure that the robot is able to learn from the human with whom it interacts. It is in the relationship, and therefore, in the mutual dependence, that learning is built and improved, both from a motivational and an ethical perspective. The adaptation to different and complex contexts, both environmental and cultural, can be explained in the same way: the interaction with the human makes the robot understand what is appropriate to do in a given situation, even in the face of unexpected changes (Redfield and Seto 2017; Murphy 2019). By doing so, the artificial agent makes sense of what happens in its context (Lawless et al. 2019): in social cooperation, the robot expands its representation of the situation to act consistently with the general goals entrusted to it. At first, this autonomous artificial agent can be quite unpredictable—though not necessarily harmful—but by continuing to learn, and therefore to interact socially, human supervision can be gradually reduced, making the robot more and more independent, albeit never completely. The human presence remains, if only for learning, but it fades in control. In interdependence, the artificial agent gradually acquires certain characteristics of autonomy that allow it to better help the human operator or user.

[11] This shared authority, in turn, seems to entail the possibility of a distributed morality (Floridi 2013).

6 Conclusion: artificial autonomy and human automation

It is time to take up our key question, which has remained in the background: do we really want autonomous artificial agents? Why? From what has been said so far, not only an autonomous social robot, being interdependent, could prove versatile and act in unstructured contexts, adapting its decisions to new and unexpected situations, assisting the human in various tasks, to the point of carrying them out directly, that is, independently. An autonomous artificial agent, placed in a social context of mutual interaction and dependence, could also learn what is normatively adequate to do or avoid. The functionality of such an artefact would be maximum, as it would guarantee efficiency, safety, reliability, and moral respect. For 'machine ethics', the goal of building an autonomous robot also lies in considering interdependence, a dimension capable of ensuring a truly trustworthy and moral robot. If this trust were to materialize, progressively more tasks and decisions would be delegated to the autonomous artificial agent, making it somehow more and more responsible.

Shifting the attention to the other side of the coin, represented by 'roboethics', it is now a question of considering the potential ethical-anthropological implications of the interactions between autonomous artificial agents and humans. In this last part, the emphasis will be placed on the human side of the interface. If robots gradually become autonomous, i.e. capable of performing (perhaps with more precision) those actions that are usually considered delicate and sensitive, and therefore, we start talking about artificial responsibility—an inevitable corollary of artificial autonomy—how does human morality change? Why do we want to endow these artificial agents with autonomy?

Roboethics—a term coined by Veruggio (2005)—deals with how humans should build, use, and deal with robots (Veruggio and Operto 2008). It is an applied ethical reflection that wants to inspire a moral development and employment of robotics (Tzafestas 2016). Among the fundamental interests of roboethics are the social and ethical implications of artificial agents: what is the impact of these technologies on the interactions with and between human beings? If machine ethics is open to the idea that robots can be considered moral subjects, roboethics states that moral responsibility always rests with human beings: in this perspective, robots cannot be ethical agents, at least not entirely. The former is interested in making machines moral, the latter in making the interaction between humans and robots moral. The distinction, however, is not clear-cut. Beyond the differences in approach and perspective, there is a common point: the relevance of the relationship between biological and artificial agents. Machine ethics and roboethics are not that distant if we consider social robotics. In both cases, interaction with the human is always involved: on one hand, it is a question of making a robot moral so that it can behave adequately in society (an ability that can be achieved in the social context itself); on the other, it is a question of investigating the anthropological consequences of this same relationship.

Let us deepen the notion of artificial responsibility. What essentially seems to define us, as humans, is autonomy. Kant calls it "the property of the will by which it is a law to itself" (Kant 1997: vol. 4: p. 440). This characteristic establishes "the dignity of human nature" (Kant 1997: vol. 4: p. 436), that is, human freedom and morality. What happens when robots are granted some kind of autonomy, however, artificial? How does the anthropological understanding of our moral character change? If artificial agents are increasingly more autonomous, are humans progressively more automatic, i.e. heteronomous?

While this latter question is purposely extravagant, and while artificial agents capable of autonomously acting and helping humans in different and multi-cultural social contexts could prove to be advantageous and thus desirable, we should not ignore the actual and potential consequences of artificial autonomy on human autonomy and morality. Why, from a moral and not functional perspective, we are granting some degrees of autonomy to artificial agents?

Maybe because we are trying "to escape precisely from autonomy, which, by making us free and moral, makes us, at the same time, potentially culpable and deserving of punishment" (Chiodo 2021: p. 3). Morality and freedom, indeed, implicate responsibility, the burdensome fact of being potentially chargeable as moral entities: if we are free and have moral agency, we can act accordingly to our self-given law, but this requires us to eventually account for what we decide to do or say. Autonomy implies dignity, i.e. freedom and morality, but these, in turn, entail responsibility: we have to explain and give reasons for what we do, possibly undergoing blame and punishment for our moral guilt. The point here is not whether robots are capable of such accountability (this topic would require many more pages). Rather, we need to reflect on the fact that autonomous artificial agents could come to act and make decisions for us, replacing our autonomy. However exaggerated and notional, this scenario has to be considered in all its ethical scope.

The technological delegation, that is, the fact of entrusting more and more decisions and activities to autonomous artificial agents, seems to be aimed at a double form of freedom: more free time, but also less responsibility for what, after all, is not our fault. In the total delegation (which, ultimately, is an extremization of interdependence), the principle of action is heteronomous and, therefore, it does not belong to us in terms of accountability. If a robot is autonomous (here

understood as independent of humans, the necessary and direct meaning of total delegation), it can decide to do something and how to do it (even just the intermediate ends): if something goes wrong, we no longer have faults, we can no longer be morally blamed or punished. The paradox is that we seem to free ourselves from our freedom by transferring responsibility to artificial scapegoats: "we seem to trade our autonomy for our freedom from individual responsibility" (Chiodo 2021: p. 5). The typical assertions of this vicious circle, which are already been heard today when something does not work or did not go as it should have been, are 'it is not my fault' or 'it is not my responsibility'.[12]

This ethical-ontological shift, in addition to being an indication of a "radical form of anarchism" (there is no longer *nomos*, therefore, no *arché*, as the internal law is removed and no free principle guides action and thinking), risks plunging us into a very original form of totalitarianism in which heteronomy and contingency are the masters. Whenever this exchange of moral prerogatives takes place, we endanger "the very core of our identity as it has been thought of in the Western culture for millennia, i.e. as rational and moral decision-makers—as autonomous humans" (Chiodo 2021: pp. 6–9).

From an ethical-anthropological viewpoint, this picture is not too promising, despite its social functionality. Here then, in the face of these conceptual concerns (which are never entirely theoretical), it seems necessary to reaffirm the value of the category we have discussed at length: interdependence. In addition to founding artificial autonomy, with all its present limitations, this concept appears to provide a confident perspective in terms of human freedom. If social robotics is implemented with this capacity, artificial entities will never be our substitutes but our respectable and respectful counterparts. In the reciprocal and mutually dependent interaction, biological and synthetic bounds and vulnerabilities will compensate each other to achieve not only more efficient but also more ethical actions. In the interdependence capable of not falling into the convenience of total delegation, we will find a dimension of mutual aid that will allow mankind to cultivate even more deeply the ability to evolve culturally, socially, and morally.

Acknowledgements We would like to acknowledge the valuable comments and feedback of the reviewers at AI and Society.

Author contributions Both authors contributed to the study conception and design. Conceptualization: FP and VGS. Writing and original draft preparation: FP (all paragraphs except 4.1). Review and editing: VGS and FP. Supervision: VGS. All authors read and approved the final manuscript.

[12] This point also concerns the biases expanded by algorithms and the consequent demands for transparency. On this see: https://www.mathwashing.com/.

Funding Not applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abràmoff MD, Tobey D, Char DS (2020) Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. Am J Ophthalmol 214:134–142

Alaieri F, Vellino A (2016) Ethical decision making in robots: autonomy, trust and responsibility. In: Agah A, Cabibihan JJ, Howard A, Salichs M, He H (eds) Social robotics. ICSR 2016. Lecture Notes in Computer Science, vol 9979. Springer, Cham, pp 159–168

Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics New Inf Technol 7:149–155

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26

Anderson M, Anderson SL (2010) Robot be good. Sci Am 303(4):72–77

Arkin RC, Ulam P, Wagner AR (2012) Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE 100(3):571–589

Baldassarre G, Mirolli M (eds) (2013) Intrinsically motivated learning in natural and artificial systems. Springer, Berlin

Baldassarre G, Lord W, Granato G, Santucci VG (2019) An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. Front Neurorobotics 13:45

Baranes A, Oudeyer PY (2013) Active learning of inverse models with intrinsically motivated goal exploration in robots. Robot Auton Syst 61(1):49–73

Bartneck C, Lütge C, Wagner A, Welsh S (2021) An introduction to ethics in robotics and AI. Springer, Cham

Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. Discrete Event Dyn Syst 13(1):41–77

Bellemare M, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R (2016) Unifying count-based exploration and intrinsic motivation. Adv Neural Inf Process Syst 29:1471–1479

Blaes S, Vlastelica Pogancic M, Zhu JJ, Martius G (2020) Control what you can: intrinsically motivated task-planning agent. In: 33rd

Conference on neural information processing systems (NeurIPS 2019), Curran Associates Inc., pp 12520–12531
Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford
Braitenberg V (1986) Vehicles: experiments in synthetic psychology. The MIT Press, Cambridge
Brooks RA (1991) Intelligence without representation. Artif Intell 47(1–3):139–159
Cangelosi A, Schlesinger M (2018) From babies to robots: the contribution of developmental robotics to developmental psychology. Child Dev Perspect 12(3):183–188
Castelfranchi C, Falcone R (2003) From automaticity to autonomy: the frontier of artificial agents. In: Hexmoor H, Castelfranchi C, Falcone R (eds) Agent autonomy. Kluwer Academic Publishers, Dordrecht, pp 103–136
Chiodo S (2021) Human autonomy, technological automation (and reverse). AI Soc. https://doi.org/10.1007/s00146-021-01149-5
Coeckelbergh M (2006) Regulation or responsibility? Autonomy, moral imagination, and engineering. Sci Technol Hum Values 31(3):237–260
Department of Defense, Defense Science Board (2012) Task force report: the role of autonomy in DoD systems. Office of the Secretary of Defense, Washington, DC. https://fas.org/irp/agency/dod/dsb/autonomy.pdf. Accessed 20 July 2021
Der R, Martius G (2012) The playful machine: theoretical foundation and practical realization of self-organizing robots, vol 15. Springer, Berlin, Heidelberg
Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Ethics Inf Technol 20(1):1–3
Duminy N, Nguyen SM, Zhu J, Duhaut D, Kerdreux J (2021) intrinsically motivated open-ended multi-task learning using transfer learning to discover task hierarchy. Appl Sci 11(3):975
Dupré J (2001) Human nature and the limits of science. Oxford University Press, New York
Floridi L (2013) Distributed morality in an information society. Sci Eng Ethics 19:727–743
Galdon F, Hall A, Wang SJ (2021) Designing trust in highly automated virtual assistants: a taxonomy of levels of autonomy. In: Dingli A, Haddod F, Klüver C (eds) Artificial intelligence in industry 4.0: a collection of innovative research case-studies that are reworking the way we look at industry 4.0 thanks to artificial intelligence. Springer, Cham, pp 199–211
Gips J (1995) Towards the ethical robot. In: Ford K, Glymour C, Hayes P (eds) Android epistemology. The MIT Press, Cambridge, pp 243–252
Grinbaum A, Chatila R, Devillers L, Ganascia JG, Tessier C, Dauchet M (2017) Ethics in robotics research: CERNA mission and context. IEEE Robot Autom Mag 24(3):139–145
Groover MP (2016) Automation, production systems, and computer-integrated manufacturing. Pearson Education, India
Hart S, Grupen R (2013) Intrinsically motivated affordance discovery and modelling. In: Baldassarre G, Mirolli M (eds) Intrinsically motivated learning in natural and artificial systems. Springer, Berlin, pp 279–300
Huang H, Pavek K, Novak B, Albus J, Messina E (2005a) A framework for autonomy levels for unmanned systems (ALFUS). In: Proceedings of the AUVSI's unmanned systems North America 2005, Baltimore, MD, pp 1–9
Huang H, Pavek K, Albus J, Messina E (2005b) Autonomy levels for unmanned systems (ALFUS) framework: an update. In: 2005 SPIE defense and security symposium, Orlando, FL, pp 1–10
Huet C, Mastroddi F (2016) Autonomy for underwater robots – a European perspective. Auton Robot 40:1113–1118
Kant I (1997) Groundwork of the metaphysics of morals (trans, edited by M. Gregor). Cambridge University Press, Cambridge (original work published in 1785)
Kohonen T (1990) The self-organizing map. Proc IEEE 78(9):1464–1480
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
Lawless WF, Sofge D (2017) Evaluations: autonomy and artificial intelligence: a threat or savior? In: Lawless WF, Mittu R, Sofge D, Russell S (eds) Autonomy and artificial intelligence: a threat or savior? Springer, Cham, pp 295–316
Lawless WF, Mittu R, Sofge D, Hiatt L (2019) Artificial intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI. AI Mag 40(3):5–13
Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137
Lo Presti P (2020) Ethical consequence of autonomous AI. Challenges for empiricist and rationalist philosophy of mind. Humana Mente J Philos Stud 13(37):19–39
Lungarella M (2007) Developmental robotics. Scholarpedia 2(8):3104. http://www.scholarpedia.org/article/Developmental_robotics. Accessed 20 July 2021
Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connect Sci 15(4):151–190
Malle BF, Scheutz M (2014) Moral competence in social robots. In: 2014 IEEE international symposium on ethics in science, technology and engineering, pp 1–6
Manoury A, Nguyen SM, Buche C (2019) Hierarchical affordance discovery using intrinsic motivation. In: Proceedings of the 7th international conference on human-agent interaction, pp 186–193
Muehlhauser L, Helm L (2012) The singularity and machine ethics. In: Eden AH, Moor JH, Soraker JH, Steinhart E (eds) Singularity hypotheses: a scientific and philosophical assessment. Springer, Berlin, pp 101–126
Murphy R (2019) Introduction to AI robotics. The MIT Press, Cambridge
Nof SY (ed) (1999) Handbook of industrial robotics. John Wiley & Sons, New Jersey
Oudeyer PY, Kaplan F (2009) What is intrinsic motivation? A typology of computational approaches. Front Neurorobotics 1:6
Oudeyer PY, Baranes A, Kaplan F (2013) Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In: Baldassarre G, Mirolli M (eds) Intrinsically motivated learning in natural and artificial systems. Springer, Berlin, pp 303–365
Powers TM (2011) Prospects for a Kantian machine. In: Anderson M, Anderson SL (eds) Machine ethics. Oxford University Press, New York, pp 464–475
Redfield SA, Seto ML (2017) Verification challenges for autonomous systems. In: Lawless WF, Mittu R, Sofge D, Russell S (eds) Autonomy and artificial intelligence: a threat or savior? Springer, Cham, pp 103–127
Reynolds N (1993) Ethos as location: new sites for understanding discursive authority. Rhetor Rev 11(2):325–338
Romero A, Bellas F, Becerra JA, Duro RJ (2020) Motivation as a tool for designing lifelong learning robots. Integr Comput Aided Eng 27(4):353–372
Ryan RM, Deci EL (2000) Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 25(1):54–67
Santos-Lang C (2012) Ethics for artificial intelligences. Version 3. https://santoslang.wordpress.com/article/ethics-for-artificial-intelligences-3iue30fi4gfq9-1/. Accessed 20 July 2021
Santucci VG, Baldassarre G, Mirolli M (2016) Grail: a goal-discovering robotic architecture for intrinsically-motivated learning. IEEE Trans Cogn Dev Syst 8(3):214–231
Santucci VG, Baldassarre G, Cartoni E (2019) Autonomous reinforcement learning of multiple interrelated tasks. In: 2019 Joint IEEE

9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob), pp 221–227
Santucci VG, Oudeyer PY, Barto A, Baldassarre G (2020) Intrinsically motivated open-ended learning in autonomous robots. Front Neurorobotics 13:115
Schillaci G, Pico Villalpando A, Hafner VV, Hanappe P, Colliaux D, Wintz T (2021) Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces. Adapt Behav 29(6):549–566
Schmidhuber J (2010) Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans Auton Ment Dev 2(3):230–247
Shanahan M (2015) The technological singularity. The MIT Press, Cambridge
Singh S, Barto AG, Chentanez N (2004) Intrinsically motivated reinforcement learning. In: Proceedings of the 17th international conference on neural information processing systems, pp 1281–1288
Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. The MIT Press, Cambridge
Tessier C (2017) Robots autonomy: some technical issues. In: Lawless WF, Mittu R, Sofge D, Russell S (eds) Autonomy and artificial intelligence: a threat or savior? Springer, Cham, pp 179–194
Totschnig W (2020) Fully autonomous AI. Sci Eng Ethics 26(4):2473–2485
Truszkowski W, Hallock H, Rouff C, Karlin J, Rash J, Hinchey M, Sterritt R (2010) Autonomous and autonomic systems with applications to NASA intelligent spacecraft operations and exploration systems. Springer, London
Tzafestas SG (2016) Roboethics. A navigating overview. Springer International Publishing AG, Berlin
Van Wynsberghe A (2021) Social robots and the risks to reciprocity. AI Soc. https://doi.org/10.1007/s00146-021-01207-y
Veruggio G (2005) The birth of roboethics. In: Proceedings of IEEE international conference on robotics and automation (ICRA 2005): workshop on robo-ethics, Barcelona, pp 1–4
Veruggio G, Operto F (2008) Roboethics: social and ethical implications of robotics. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer, Berlin, pp 1499–1524
Wallach W (2008) Implementing moral decision making faculties in computers and robots. AI Soc 22:463–475
Wallach W, Asaro PM (eds) (2017) Machine ethics and robot ethics. Routledge, London
Wallach W, Allen C, Smit I (2008) Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI Soc 22:565–582
Wallach W, Franklin S, Allen C (2010) A conceptual and computational model of moral decision making in human and artificial agents. Top Cogn Sci 2:454–485
Washington R, Golden K, Bresina J, Smith DE, Anderson C, Smith T (1999) Autonomous rovers for Mars exploration. In: 1999 IEEE aerospace conference proceedings, vol 1, pp 237–251
White RW (1959) Motivation reconsidered: the concept of competence. Psychol Rev 66(5):297
Yang GZ, Cambias J, Cleary K, Daimler E, Drake J, Dupont PE, Hata N, Kazanzides P, Martel S, Patel RV, Santos VJ, Taylor RH (2017) Medical robotics – regulatory, ethical, and legal considerations for increasing levels of autonomy. Sci Robot 2(4):eaam8638. https://doi.org/10.1126/scirobotics.aam8638
Yip M, Das N (2018) Robot autonomy for surgery. In: The encyclopedia of medical robotics, pp 281–313
Zeigler BP (1990) High autonomy systems: concepts and models. In: IEEE proceedings. AI, simulation and planning in high autonomy systems, pp 2–7