15

Artificial Intelligence and New Paradigms of Human Decision Making: Towards a New Idea of Humanity?

Antonino Rotolo

1. Introduction

The human digital transformation is a disruptive transition of humanity to a new era, it is in fact 'the fourth revolution' in which humans significantly live and interact in the 'infosphere' (Floridi 2014). The infosphere is the reconfiguration of the world and human social interaction in which we live online, with the impossibility of clearly distinguishing online from offline. Therefore, the digital is not just a medium, but it is a dimension of human existence, within which to act, know and forge relationships.

The impact of the digital transformation works with a 'double direction of fit':

world → digital
digital → world

The digital transformation is thus reshaping social dynamics.

Within the transition to the digital, the emergence of Artificial Intelligence (AI) raises several questions regarding:

- how a humanistic approach to AI can make an epistemological, ethical, legal and cultural impact in our contemporary world (*AI is designed according to a new idea of humanity*);
- how AI can impact on traditional humanistic issues and on the humanities understood as a field of knowledge (*AI transforms and reshapes the idea of humanity*).

We can predict a near future in which human societies will make a pervasive use of artificial autonomous systems or agents, such as self-driving vehicles, drones, a large variety of AI software platforms such as everyday or professional applications of advanced large language models, intelligent digital personal assistants, home hubs, and a multitude of industrial applications (cf. Corea 2018; Floridi 2023: 31–54). With autonomy, awareness is also likely to come. Beside the dimension of self-awareness, another dimension of awareness will be increasingly relevant: this is norm-awareness, by which we mean an agent's knowledge and adoption of norms governing its behaviour. In particular, this dimension concerns instructions given to the AI systems by humans, but also legal rules.

To be clear, we are not arguing that AI will be like humans, nor that it will lead to a dystopian world where moral dilemmas pervade our life (cf. Floridi 2019). However, we believe that several concepts pertaining to the idea of humanity can be reshaped (in both directions: see above) according to the AI paradigm turn.

2. Autonomy

An autonomous artificial agent will be delegated the more or less autonomous execution of a various task; namely, which will have to be a distributed plan including aspects delegated to humans, aspects delegated to machines, and aspects that require the integrated human and mechanical agency (cf. Castelfranchi and Falcone 1998).

Autonomy in a delegated agent may increase along various dimensions: the more open the delegation (the less specified the task is), or the more control actions are given up or delegated to the agent, or the more delegated decisions (discretion) are entrusted to the agent. Accordingly, an autonomous agent can become the addressee of a more and more 'open delegation' (cf. ibid.), in the sense that it will be assigned tasks the execution of which are less and less specified. This means that an autonomous agent delegated to take care of a given task has to choose between different possible recipes; or it has to build a new plan, or to adapt abstract or previous plans to new situations; it has to find additional (local and updated) information; it has to solve a problem (not just to execute a function, an action, or implement a recipe); sometimes it has to exploit its 'expertise'. An agent involved in a multi-agent plan must also understand its role in the plan, as an autonomous executor of the plan together with other similarly located executors of the plan, which may also be in charge of coordination functions.

Autonomous agents are supposed to use their knowledge, intelligence and ability, and to have some autonomy of decision. This is not a constraint on the use of such an agent, a defect we may want to remedy, this is exactly the reason why we are using such an agent, since humans are unable, or are unavailable or are too expensive to execute such cognitive tasks. We delegate to them cognitive tasks pertaining both to epistemic cognition (knowing how things are) and to practical cognition (knowing what should be done), since we are unable to provide the agent with complete prior knowledge, and we choose to rely on the agent's cognitive efforts, on its percepts, the information it extracts from them and its subsequent determinations.

3. Awareness and (Human–Machine) Communication

An autonomous system, entrusted with an open delegation, will need the ability to consciously perceive itself and its position in basic social interactions, as a collaborator in distributed common plans with its human and artificial 'fellows'. In fact, building new plans requires the agent to distinguish itself from the environment and, in particular, participating in multi-agent plans requires the agent to recognise other agents as different from itself.

The more systems are endowed, the more difficult it is to understand and anticipate their behaviour on the basis of the working of their internal mechanisms, and, in particular, considering the programming instructions implemented into them. When dealing with autonomous systems, a useful option is to adopt Dennett's 'intentional stance' (Dennett 1987); namely, the strategy of interpreting the behaviour of such entities through the mentalistic notions typically applied to human agents, such as knowledge, belief and intention. Here we see an interesting development of how such mentalistic notions, if imported from the human domain into the artificial one, can reshape the idea of cognitive subject: the concept of humanity will change accordingly.

Notice that one could think that autonomy and intentional stance, if combined, may lead one to attribute full personhood (for example, in the moral or legal sense) to AI subjects. This conclusion has been advocated by several philosophers. Indeed, if an

artificial agent can autonomously decide and is self-aware, personhood is around the corner. However, if robust approaches to self-awareness are employed, such a conclusion looks unreasonable. In particular, the *embodied-cognition paradigm* argues that the human mind is significantly determined by the form of the human body. Hence, we may see the self as a dynamic concept emerging from a constructive process where a subject is first of all embodied and, as such, it starts with reifying itself and with ascribing cognitive dimensions to itself. It is worth noting that neurosciences more often insist that cognition is embodied, so *self-awareness, too, requires a necessary move from the body to the mind dimension* (cf. Damàsio 1999). From this perspective, full personhood can hardly be attributed to AI systems.

A more reasonable (and simpler) perspective on the above issue is the following. An artificial agent will be directed by assigning it tasks to be achieved, and its behaviour will be monitored by assessing whether it is an appropriate or rational way of achieving such tasks in the context in which they are executed. Since the appropriateness (and the legality) of the performance of such actions involves taking into account the impact of the system's activity not only on the goal assigned to it, but also on other values which come to be at stake, under unpredictable circumstances, it remains to be seen when and to what extent tasks whose adequate performance involves these evaluations can be entrusted to machines.

According to several studies in the area of cognitive systems engineering, when in a system the automation has taken over more or less completely, humans become controllers of automated systems rather than operators. These systems exercise cognitive functions, they acquire information from the environment, process it, and use the knowledge so obtained to achieve the goals assigned to them, as specified by their users. It has been observed that when one or several operators and one or several automated support systems interact together for the fulfilment of a task, it would be better to describe humans and technology not as two interacting 'components' but as constituting a joint (cognitive) system. The term 'joint cognitive system' means that control is accomplished by an ensemble of cognitive systems and (physical and social) artefacts that exhibit goal-directed behaviour. Several studies also describe these *fusions between humans and machines as 'hybrids'* (cf. Kamar 2016). In hybrids, the participating individual or collective actors are not acting for themselves but are acting for the hybrid as an emerging unit, the association between humans and non-humans. They do so in the same way that managers are not acting on their own behalf but are 'agents' or 'actants' representing their 'principal', which is the corporation as a social system. In these cases, agency does not pertain only to humans or to machines, but to the hybrid itself, so that human–machine interaction and trust play a decisive role in assessing their joint behaviour. In addition, we need to understand how such a hybrid entity can be governed, what norms should apply to the human component and what constraints should conversely address the machine component, where the interaction between the two should be designed to minimise mistakes and carelessness, control aggressiveness, and implement legal and moral constraints.

4. Responsibility

Finally, high autonomy can make it harder to identify responsibilities for wrongdoing as well as any type of damages caused (cf. Corrales et al. 2018; Pagallo 2013). This holds especially true when these effects are unexpected and do not result from a system failure,

since the autonomous system itself autonomously learned how to achieve the effect. In this sense, the resulting consequence is in fact unexpected (and, possibly, beyond any predictive capability) only from the perspective of the system designer, not from the perspective of the AI system, which seems to possess the required subjective states.

In general, the ascription of legal responsibilities is an open problem when autonomous systems are considered. For example, can an autonomous vehicle be held responsible for personal injury? Are we ready to accept the idea that AI in such a vehicle would choose to avoid a child on the road, saving that child, and kill a dog by hitting it? What if traffic conditions force us, for safety reasons, to violate speed limits? The discussion on these issues has contrasted optimistic and critical or even catastrophic views in recent years. In particular, in the contemporary debate there are those who believe that AI poses unprecedented ethical problems that are configured as real dilemmas; for example, as new variants of the well-known scenario of the Trolley Problem: if it is so, the task of those who develop AI is to know *how to solve these dilemmas*. Others, however, believe that this analysis handles the problem with too abstract a perspective: the scenarios described by the aforementioned ethical dilemmas are not realistic if the use of AI takes place in suitable contexts: if so, the problem lies not in solving the dilemmas, but in *creating the conditions for these scenarios to be avoided*. In this second perspective, consider Luciano Floridi's argument on the 'enveloped AI' (cf. Floridi 2019): we envelop environments around AI systems to fit and exploit their capacities in a safe and realistic way.

In the context of criminal law, there are significant conceptual and philosophical difficulties associated with attributing responsibility, and, moreover, it is not clear whether it makes sense to punish non-human entities, especially artificial machines. The difficulty raised is that future AI systems, because of their fully autonomous potential, may become capable of causing damage or inflicting violence that falls within the categories of crimes or wrongdoings. Yet, we assume that an AI system is incapable of bearing effective responsibility for the harm it causes, despite being in an analogous situation to a direct perpetrator, because it is not a natural person and so it cannot be punished in a significant way according to our law. (What about the retributive and deterrent function of the criminal sanction?) This possibility, for example, for robotic weapons systems to replace the direct perpetrator of a crime may cause the structure of responsibility to collapse.

In general, however, while with full personhood the combination of intent and causation may jointly lead to responsibility, this hardly applies to current AI conceptions. Consider, for instance, that in several legislative contexts the fact that a subject has procured a person for a criminal offence who is not indictable or is not punishable (e.g., minors) makes the subject liable for that offence, and an increased penalty is applied. Suppose that Mr Smith induces a robot to threaten Mr Jones, and the robot is bound to that goal (to threaten Mr Jones) but is equipped with autonomy in achieving it. It should be noted that the above provision covers cases where the procured person is not legally capable, but robots, though intelligent, are not indictable and the principle of legality in criminal law does not allow the provision to be applied by analogy to the case in which the crime is committed by a robot.

While the individual criminal responsibility keystone has been removed, it has been argued that a system of distributed responsibility may compensate for the resultant structural weakness. It has been claimed that the lack of individual criminal responsibility can instead be solved by imposing command or superior responsibility.

5. Thought Experiments from AI and Literature

The idea of autonomous systems has elicited a lot of interest from science fiction writers, which has focused on different aspects of awareness. Interestingly, in such contributions the idea of autonomy is coupled with the idea of awareness, which has some interesting and specific implications. Particularly effective examples in literature depict dystopian scenarios where AI is at the core of the development of autonomous weapon systems.

First of all, obviously we must recall the work of Isaac Asimov. In his work the idea of norm-awareness is indeed central. Asimov's scenery is dominated by the famous three laws of robotics: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

The norm-awareness in Asimov's robots is coupled with the strongest degree of norm-motivation; such robots are incapable of acting in contravention to the three laws, and in particular with the first law. The mere fact of damaging a human inadvertently would cause the robot to break down, damaging its circuits permanently. However, the idea of autonomous (killer) weapons (as opposed to anti-materiel weapons) occurs in different ways in Asimov's stories as humans conceive and implement the possibility of building robots able to deviate from the laws or when the robots themselves take the initiative to do so. The possibility that harm is done to an individual for the long-term benefit of humanity is conceived by a benevolent robot in the novel *Robots and Empire*. This robot conceives a higher-level law, the Zeroth Law, reading: 'A robot may not harm humanity, or, by inaction, allow humanity to come to harm.' So it seems that Asimov's benevolent robots may want to engage in humanitarian wars, to use violence and even lethal force to protect human rights. However, even the most capable of the robots has to admit that he was in the end unable to use the Zeroth Law as an authorisation for killing, since 'Injury to a person can be estimated and judged. Humanity is an abstraction.'

The concrete possibility of using robots against human beings takes two possible directions in Asimov's universe. First of all, there is the possibility that robots restrict the domain of the category of humanity, limiting it to a set of humans, so that the robots may harm certain individuals while complying with the law, since they interpret the term 'human' as not applying to such individuals. The other possibility is to have robotic weapons fully free from the law, a possibility which is developed by a roboticist in the book *The Naked Sun*. This describes a society in which each human uses a number of robots. A scientist starts to develop autonomous weapons (war starships) free from the three laws, but his society, so dependent on robots, rejects this idea, fearful of what may happen if the robots, once free from the laws, revolt against their human masters.

The idea that autonomous weapons able to kill may be used in a war scenario, to engage in atrocities, which increase as their autonomy increases, freeing themselves not only from norms but also from the link to their principals, is developed by Philip K. Dick in the novel *Second Variety*, where intelligent killer weapons developed for global warfare acquire the ability to construct and perfect themselves, become more and more similar to humans, and end up wiping out humanity. Interestingly each party to the war builds such weapons with the awareness of the enemy (so that one's autonomous weapons only attack one's enemy), but as new technologies are developed to trick the weapons into perceiving or not perceiving others as enemies, this awareness becomes so sophisticated that the choice of the enemy falls upon the weapons themselves, who engage in sweeping out humanity.

That self-awareness may finally lead a weapon system to adopt its own purposes, which may be different and possibly opposite to those of its users, is also the basis of a number of popular movies, from the *Terminator* series to *I, Robot*. In such movies, self-awareness in an autonomous weapon system leads the system to endorse a will to persist (to survive), which becomes a will to power (to have and control the resources to survive and expand), which in its turn may lead to a conflict with the human principals when they try to contain or remove such powers.

A different picture emerges from another famous science fiction work; namely, Stanisław Lem's essay 'The Upside-Down Revolution' in his book *One Human Minute*, which describes a future where war is delegated to synthetic insects (synsects), namely 'ceramic microcrustacea, titanium annelids, and flying pseudo-hymenoptera with nerve centres made of arsenic compounds and with stingers of heavy, fissionable elements'. Such microsoldiers are endowed with swarm intelligence, and can build themselves: they are designed, combat-tested and sent to be mass-produced by 'construction battalions' of non-living microdesigners. The artificial insects establish pervasive surveillance (thanks to them anything can be a covert agent: a nail in the wall, a laundry detergent, etc.) and they dominate the world, and even merge with nature, to produce apparently natural disasters. Though the novel does not provide details on how the swarm intelligence may emerge, it does provide for an anticipation of the possible development of autonomous weapons, where awareness resides at the collective level in a set of self-coordinating agents.

A fourth challenging perspective, where autonomy and awareness are maximised in an artificial system, is provided in the *Culture* series by Iain M. Banks, which describes a most advanced society, governed by super-powerful artificial intelligences, called Minds. Minds in charge of warships have a fully developed psychology, including suprahuman cognitive capacities, moral awareness and moral sentiments. The novel *Look to Windward* includes the suicide of the Mind of a starship, due to its incapacity to sustain the sense of guilt resulting from its behaviour in war, where it had to cause the death of a large number of people.

6. Recommendation for the New Humanities

In conclusion, we can predict a near future in which human societies will significatively live online or, more generally, in the digital world. We need new humanities to devise comprehensive epistemological, ethical, legal and cultural models for our contemporary world. Humanities are thus expected to reshape the idea of humanity. The above discussion leads to the following recommendation:

> To promote a reflective and generative role of humanities in the digital era, i.e., humanities as a way to change social paradigms, to address global challenges and reorient them by envisioning long- and short-term innovation.

Acknowledgement

Parts of this chapter elaborate ideas from Bhuta et al. (2015).

References

Bhuta, N., A. Rotolo and G. Sartor (2015), 'Awareness and Responsibility in Autonomous Weapons Systems'. In J. Pitt (ed.), *The Computer After Me: Awareness and Self-Awareness in Autonomic Systems*. London: Imperial College Press.

Castelfranchi, C. and R. Falcone (1998), 'Towards a Theory of Delegation for Agent-based Systems'. *Robotics and Autonomous Systems* 24: 141–57.

Corea, F. (2018), 'AI Knowledge Map: How to Classify AI Technologies – A Sketch of a New AI Technology Landscape'. *Medium – Artificial Intelligence*. https://medium.com/@Francesco_AI/ai- knowledge-map-how-to-classify-ai-technologies-6c073b969020

Corrales, M., M. Fenwick and N. Forgò (2018). *Robotics, AI and the Future of Law*. Berlin-New York: Springer.

Damàsio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt.

Dennett, D. (1987), *The Intentional Stance*. Cambridge, MA: The MIT Press.

Floridi, L. (2014), *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press.

Floridi, L. (2019), 'What the Near Future of Artificial Intelligence Could Be'. *Philosophy & Technology* 32: 1–15.

Floridi, L. (2023), *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Kamar, E. (2016), 'Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence'. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence 2016*. Menlo Park, CA: AAAI Press.

Pagallo, U. (2013), *The Laws of Robots: Crimes, Contracts, and Torts*. Berlin and New York: Springer.