Artificial Intelligence: "Machines of loving grace" or "Tools for conviviality"?

Wolfgang Hofkirchner
GSIS The Institute for a Global Sustainable Information Society, Vienna, Austria
https://gsis.at

In 1967, when being Poet-in-Residence at the California Institute of Technology, **U.S. writer Richard Brautigan** published a poetry collection. In one poem with the title **"All Watched Over by Machines of Loving Grace"** he fancies "a cybernetic ecology |where we are free of our labors | and joined back to nature, | returned to our mammal | brothers and sisters, | and all watched over | by machines of loving grace" (quoted in Madrigal 2011). This is a technotopian imaginary of a **benevolent technocracy safeguarding an all-out harmony**.

Six years later, in the aftermath of the hippie movement and the student revolts, **Austrian born writer Ivan Illich**, working as parish priest, university rector, or professor in the field of Science–Technology–Society (Penn State University), and commuting between Mexico, the U.S., and Germany, published a book with the title **"Tools for conviviality**". In this book, he submits "the concept of a multidimensional balance of human life which can serve as a framework for evaluating man's relation to his tools. In each of several dimensions of this balance it is possible to identify a natural scale" (1973, x). "Once these limits are recognized, it becomes possible to articulate the triadic relationship between, persons, tools, and a new collectivity. *Such a Society, in which modern technologies serve politically interrelated individuals rather than managers, I will call 'convivial.'* […] *I have chosen 'convivial' as a technical term to designate a modern society of responsibly limited tools*" (xii). 'Convivial' has Latin origins and means the quality of living together in the manner of dining together (*convivor*) of hosts (*convivatores*) and guests (*convivae*) at common feasts (*convivia*). It shall not mean "tipsy jollyness" but "*eutrapelia* (or graceful playfulness)" – going back to one of the virtues of Aristotelian ethics elaborated by Thomas Aquinas, which is associated with "friendship or joyfulness" (xiii). During the last decade, convivialism has become a social movement, initiated by French intellectuals around Serge Latouche, Edgar Morin and more. Compared with Brautigan, Illich's imaginary of the future sketches a balanced society too, but it reverses the order of influence: a **convivial society shapes technology for conviviality**.

Both options share an optimistic picture of the future development of culture and civilisation. The question is: which of the two imaginaries is more apt to characterise desirable applications of Artificial Intelligence in the societies to come – machines of loving grace or tools for conviviality? Which is more realistic and more desireful?

Roberto Simanowski (2020), German expert in German Literature and Media Studies, is partisan of the first option. In 2018, he had already published two essays with MIT Press in *The death algorithm and other digital dilemmas* that, after significant expansion and revision, form the first two chapters in the 2020 book in German. The death algorithm is the programme that in case of imminent accidents steers the self-driving car into a target of choice. Simanowski shows with rigour **aporias and paradoxes that cannot be solved on the basis of programming vehicles according to utilitarian/consequentialist or deontological ethics, because no programme will satisfy a universal rule acceptable for all humans**.

In that context, Simanowski cites another example – the drama **"Terror"** written by German writer Ferdinand von Schirach. In that play, the theatre or TV audience sits in judgement on a fictive Major of the German military forces who, in an unauthorised act, shot down a civil airplane with 164 passengers in order to save the lives of 70.000 people in a Munich arena into which the airplane was supposed to be downed by a terrorist in control of the machine. The majority of the audiences voted for an **acquittal** of the accused Major – namely, 63 per cent of the playgoers in 2.472 performances from October 2015 until January 2020 and 87 per cent of the television viewers of the movie at German, Austrian and Swiss TV stations on 18 October 2016 (Simanowski 2020, 19-20). Such a decision would be in violation of the constitution of the German Federal Republic that forbids to offset one number of casualties against another number of casualties. Furthermore, there are different majorities in different cultures. Asian values might be different. During the same period, 11 performances out of a total of 21 in China as well as 15 out of a total of 23 in Japan resulted in **convictions**, whereas all of 8 performances in Taiwan ended with acquittals (Simanowski 2020, 40). This difference in votes holds also for other examples (like the Trolley Dilemmas) and might have to do with the **difference between individualistic and collectivistic cultures** (Ahlenius and Tännsjö 2012, quoted in Simanowski 2020, 129).

Simanowski understands that the advent of **self-driving cars worldwide is not compatible anymore with the current status of ethics**. It calls for a universal solution. But for Simanowski such a solution must be different from a universal ethics which he considers impossible. Hence his idea of Artificial Intelligence as solution – not the weak AI but a strong AI. He believes that Deep Learning might enable strong AI not only to follow decisions given by human intelligence but also to make its own decisions independently of human intelligence. And thus, **strong AI might be able to make decisions that human intelligence is still unable to make** because vested interests frustrate and cancel each other, while strong AI might be disinterested in human particularisms **and neutrally watch over humans that fulfil decisions AI would make**. By doing so – that's the hope of Simanowski – strong AI might even be able to **help humanity survive**.

However, Simanowski does not understand that the aspirations of strong AI defenders are not substantiated. They will not come true, and what could come true – irrespective of what the aspirations are – would only endanger the further development of our species on Earth. Why? Not because humanity would need to fear that AI would become a malevolent being instead of a benevolent one since AI technology is, in principle, **incapable of malevolence and benevolence**. But AI can **disrupt the autonomy of humans**.

In the course of **industrialisation**, tools have been refined so as to yield machines that entered between the object of work and the worker, not any more dependent on the energy of wind and water and animals. In the course of **informatisation**, machines have been computerised so as to yield automatons that made an even bigger distance to the worker, as some functions of the human mind were conveyed to them. As Karl Marx had foreseen, **automatisation** can lead to a social state of affairs "where we are free of our labors" (Brautigan). But automatisation need not lead to **autonomisation**. Autonomy is a term that is borrowed from the realm of humans and society and imposed on machines, as if a machine would be a person. Autonomy of automatons shall signify in engineers's speak that automatons also reach decisions. If this would be true, it would mean that functions of the human mind would be taken over by automatons such that the human mind would be

deprived of those functions: either the automaton or the human takes the decision – only one of them can take it. But the situation is different: **no automaton can really reach a decision**. A decision is a **judgement** – the result of deliberating on grounds and these grounds do not prejudice the judgement. The judgement is an act based upon the grounds as necessary presuppositions, but it is not logically derivable like a conclusion from premises. There is more to a judgement. It is an **emergent** act. Emergence means that the emergent has **another – a new – quality** compared to that from which it emerges. No machine, no automaton, no so-called autonomous technical system, can produce emergence since all underlying information processes are, ultimately, deterministic mechanisms. So, what engineers call decision in the case of a machine is, actually, a product of **mechanical determinacy devoid of deliberation of a self**. Moreover, given the complicatedness of modern machines, the products of their processing have become **unpredictable** – therefore they were mistaken as emergent by engineers – and even, practically, **not explainable** (retrodictable), though every step of the algorithmic processes follows a determined rule.

This is the reason why Simanowski's faith in strong AI saving mankind is doomed to failure. The so-called decisions of strong AI would be random and inapt. AI does not dispose of an own self. It is not agential, it is rather patient (Capurro 2012). It is not a self-organising system, it is hetero-organised, externally-organised. The *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS)* published a comprehensive document on ethically aligned design that states (2019, 41) with reference to Capurro (2012) and Hofkirchner (2011): "Of particular concern when understanding the relationship between human beings and A/IS is **the uncritically applied anthropomorphic approach** toward A/IS that many industry and policymakers are using today. This approach **erroneously blurs the distinction between moral agents and moral patients, i.e., subjects, otherwise understood as a distinction between 'natural' self-organizing systems and artificial, non-self-organizing devices**. As noted above, A/IS cannot, by definition, become autonomous in the sense that humans or living beings are autonomous. With that said, autonomy in machines, when critically defined, designates how machines act and operate independently in certain contexts through a consideration of implemented order generated by laws and rules. In this sense, A/IS can, by definition, qualify as autonomous, especially in the case of genetic algorithms and evolutionary strategies. However, **attempts to implant true morality and emotions, and thus accountability, i.e., autonomy, into A/IS blurs the distinction between agents and patients and may encourage anthropomorphic expectations of machines** by human beings when designing and interacting with A/IS." Because AI, whether weak or strong, is not capable of emergent information, it cannot act in a benevolent manner. It cannot reproduce how it would be to be human. It is totally detached from a human take at the problems at hand.

Altogether, **AI cannot be the universal solution** Simanowski is looking for, since it cannot promote **unity-through-diversity** that emerges as an intrinsic problem-solving attempt in the course of evolution, **a meta-level that supervenes the level of conflicting interests of individual partitions** and nests – **as a Third** – the level below as a Second and the parts of the Second as Firsts. Only such an emergence of a Third would be a universal (though not absolute) solution. Such a Third cannot be expected by technology *per se*. And it is for that reason that the rule of AI would devolve into a rule of technocratic dictatorship in which one arbitrary situation would be superseded by another arbitrary one – **the autonomisation of machines would cause the de-autonomisation of humans**.

For **Austrian philosopher Günther Anders** this would be another example of **"Promethean shame"**, which after Copernicus, Darwin and Freud is the **fourth blow to humanity's sense of itself**. Anders described the first historical example in his essay "Über prometheische Scham" in 1956 (Anders 1956, English translation published in 2016). US General Douglas MacArthur, "at the beginning of the Korean War proposed measures that arguably could have triggered a Third World War. […] the decision as to whether such an outcome should be risked or not was taken out of his hands. Those who removed his responsibility from him, however, […] removed the decision […] to hand it over to a *machine*", "to an '*electric brain*'" (Anders 2016, 58). The "electronic brain" opted against MacArthur's approach, but Anders's emphasis is on the fact that "the process, as such, by which this decision was reached, was at the same time **the most epoch-making defeat that humanity could have inflicted on itself. For never before had humanity degraded itself to such a degree that it entrusted judgement about the course of history, perhaps even about whether it may be or not be, to a thing**" (60-61). Anders called this development the "*climax of all possible dehumanisation*" (44), "*arrogant self-degradation*" and "*hubristic humility*" (49) – **hubris** because man believes himself to be capable of constructing superhuman machines, **humiliation** because, in order to become a master, man has to turn himself into a slave.

Having said this, the *prima facie* harmonic picture of "machines of loving grace" is not only undermined but is also in stark contrast to the second option, the claim to "tools for conviviality". It is just because strong AI will never dispose of the epistemic capacities preached again and again but will limit human autonomy instead, that **AI needs to be designed meaningfully and mindfully and needs to be responsibly limited**. The German title of Illich's book "Tools for conviviality" is "Selbstbegrenzung", which means "self-limitation".

Self-limitation has now been placed irrevocably at the top of the agenda of humanity. The age we live in, the **Anthropocene**, a term supposed by geologists, is rather a **Monetocene** – as philosopher Richard David Precht in his book on AI and the meaning of life says (Precht 2020) – or a **Capitalocene** – as physicist, Harald Lesch says (2018) – in which the interests of realising profit is the driving force that changes our planet and it changes our planet to the worse. The basic question is whether or not humanity can be saved. According to Precht, **either capitalism will be overcome or *homo sapiens* will do away with himself** (2020, 11).

Conviviality in the sense of self-limitation is a feature of emergent social systems and can be determined as the **historical-concrete shape of the social relations of commoning**, that is, relations that enable and constrain social actors to contribute together to some social good through the effort of combined productive energies as well as to be common beneficiaries when consuming that good – that relational good is a common good, it is a **commons**. This is the Third sought-for, emergent from the co-action, the interaction and the actions of actors. This Third relates the actors to each other in letting them assume the roles of the Second (*alter*) and the First (*ego*). Conviviality expresses the degree to which the commons are open to any participant of the social system. **The more the actors limit selfishness, the more the commons are open to all – including themselves**. Conviviality is a vision. It would "make the social systems inclusive through the disclosing of the enclosed commons and, by doing so, [...] warrant eudaimonia, a good life in a good society, the flourishing of happy individuals in convivial social relations" (Hofkirchner 2017, 286).

Though even disciplines like mathematics or biology evidence that *homo sapiens* excels as a race of **possible super-co-operators** (Nowak and Highfield 2011), *sapientia* requires substantial further development: co-operation for the commons is still fragmented.

The history of mankind up to the current point of social evolution shows two decisive steps. The **first step** was done when our ancestors stepped out **from animal monads into social dyads**. Those ancestors adopted, in the context of common foraging, "more complex forms of cooperative sociality" (Tomasello 2014, 31) that guaranteed the common good for the included actors. By that they started a ratchet effect that yielded ever higher complex co-operation until, in a **second step** in evolution, the social factors outbalanced biological factors and **societal triads complemented the social dyads**. The triads established a **Third** – a common culture, collective intentionality and objective morals (Tomasello 2014; 2016), all of which have since been relating "individuals to each other with respect to the common good – even if the concrete content of the common good became a matter of disputation and conflict" (Hofkirchner 2020a, 5). This Third has come in two varieties.

Tribalism was the primeval variety that appeared as stage at the dawn of societal evolution (Donati 2010). There was a rather collectivistic relational "We", myths conveyed the tradition, and means and ends of social life were not questioned. Another variety originated when tribal "We"s switched to heteronomic societies (Hofkirchner 2014, 84): the actors became self-regarding persons and their thinking became short-sighted, not taking into consideration harmful effects on other parts of the system, and the structures of the social systems have been prioritising competition on the higher levels of society while co-operation has been reserved for the lower levels; the supreme good was believed to be the private; and means and ends were decoupled insofar as means were intelligently flexibilised whereas the final end stayed as a given. This rather individualistic stage lasts until today. It deserves the name "**Idiotism**" (Curtis 2013). Etymology shows, in Greek Antiquity *idios* meant "the personal realm, that which is private, and one's own" (Curtis 2020, 12). In Curtis' view, *idios* bears also the stamp of "being enclosed". He says that "the creation of the private through the enclosure of public or commonly held resources has historically been the primary means by which property has been secured for private use" (12). By the term *idiotes*, then, a person was denoted that is concerned with his personal realm only, with his own, and not with the *res publica* or *res communis.* Nowadays, neoliberalism carries idiotism to extremes. The unfolding of conviviality has been more and more challenged by such exclusive commoning.

However, "a **third step** of anthroposociogenesis can be hypothesised. There might be a shift from collective intentionality to one that is shared universally, that is, on a planetary scale. That would be the transition to another convivial regime – **an extension of the triad to the whole of humanity, an omniad**" (Hofkirchner 2020a, 5). That step would not only complete the first two steps in anthroposociogenesis but also sublate tribalism and idiotism. It could integrate individuals without requiring their subsumption. As global citizens, they would be capacitated to **reflect their own position and the position of others from the perspective of the overall social system**. No one-sided ideology nor mythologisations would obscure a realistic, science-based and practical assessment of different paths of societal development. Not only means would be variable but also the ends would not be constants anymore; and none of them should be in force unless agreed upon in common (Hofkirchner 2014, 84-85).

AI is a technology and is a specific means for some ends. As with any technology – that is, methods, procedures, artefacts – also in the case of AI, mechanisms are designed to mediate the fulfilment of social functions. A cause is functionalised as a means so as to effect an end. **When designing technology, responsibility is taken over** in two regards. **First**, responsibility is taken over for the **functionality** of the design: does the mechanism effectively and efficiently lead to the end for which technology shall be designed, that is, is it functional? This is a matter of fact. **Second**, responsibility is taken over also for the **meaningfulness**, the **social usefulness**, of the design: does the end for which the mechanism is designed make sense, that is, does it promote a social value that is worth promoting, does it conform with a social norm that is worth conforming with? This is a matter of morals. Both the functionality and the meaningfulness of technology need to be responsibly reflected. In the time of global challenges, this reflection means that, following the vision of the good society, populated by individuals living the good life and cultivating the common good, **any technology should support the transformation of societies into a Global Sustainable Information Society** (Hofkirchner 2020a, 5-6). According to that, **AI must be a technology of supporting human intelligence focused on securing social evolution from self-inflicted breakdown**. The overall goal must be set by human decision-making and AI can help find ways for implementation, can monitor target achievement and can give cause for measurement adaptations. AI, as any technology **socially embedded**, is thus **part of a techno-social system**. "Since AI is a tool that shall afford the intelligent behaviour of actors, it shall not be given room to constrict the autonomy of actors" (Hofkirchner 2020b, 3). It is up to human beings to "rationally choose the way humans evolve themselves", as Kun Wu (2020, 3) states.

Intelligence reflects (on) the means-end relationship, in particular, the functionality of the means. However, in order to be capable to catch the meaningfulness of the end, human intelligence in this narrow sense is not enough. The global problems are man-made and need a human(e) solution. For that, **wisdom** is required. Thus, society does not only need to become an intelligent society, it needs also to become a "**wise society**" – a term put forward by Spanish sociologist Manuel Castells and other academics when working with the High-Level Expert Group of the European Commission in 1997 on the topic of a European information society for all (European Commission 1997, 16). They identified wisdom "as 'distilled' knowledge derived from experience of life, as well as from the natural sciences and from ethics and philosophy". While new ICTs had been energising economy, "these new technologies have had no such effect on the generation or acquisition of knowledge and still less on wisdom. One would hope, of course, that society would be shifting more and more towards a 'wise society', where scientifically supported data, information and knowledge would increasingly be used to make informed decisions to improve the quality of all aspects of life. Such wisdom would help to form a society that is **environmentally sustainable**, that takes the **well-being of all its members** into consideration and that values the **social and cultural aspects of life** as much as the material and economic. Our hope is that the emerging information society will develop in such a way as to advance this vision of wisdom."

So far, these hopes did not yet come true and AI seems to meet a similar fate. A new "**digital humanism**", as publicised by the *Vienna Manifesto* (2019) in the wake of Julian Nida-Rümelin and Nathalie Weidenfeld (2018), can give fresh impetus to **humanise digitisation** and it can **limit AI to convivial tools**. The phantasy of machines watching over humans boils down to **anti-humanism**. **Humanism** means convivial AI to help humanity survive and thrive.

References

Ahlenius, H., Tännsjö, T. 2012. Chinese and Westerners respond differently to the Trolly Dilemmas, in: *Journal of Cognition and Culture* 12 (3-4), 195-201.
Anders, G. 1956. *Die Antiquiertheit des Menschen: Über die Seele im Zeitalter der zweiten industriellen Revolution*. Beck, München.
Anders, G. 2016. On Promethean Shame, in: Müller, C. J., *Prometheanism: technology, digital culture and human obsolescence*, Rowman and Littlefield, London, 29-95.
Capurro, R. 2012. Toward a comparative theory of agents, in: *AI & Society*, 27(4), 479-488.
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. 1st ed. IEEE, https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html.
European Commission, Directorate-General for Employment, Industrial Relations and Social Affairs (ed.) 1997. *Building the European information society for us all, Final Policy Report of the high-level expert group*, Office for Official Publications of the European Communities, Luxembourg.
Hofkirchner, W. 2011. Does computing embrace self-organisation? in: Dodig-Crnkovic, G., Burgin, M. (eds.), Information and computation, World Scientific, Singapore.
Hofkirchner, W. 2014. The commons from a critical social systems perspective, in: *Recerca*, 14, 73-91.
Hofkirchner, W., 2017. Creating common good – the global sustainable information society as the good society, in: Archer, M.S. (ed.), *Morphogenesis and human flourishing*, Springer, Dordrecht, 277-296.
Hofkirchner, W. 2020a. A paradigm shift for the Great Bifurcation, in: *BioSystems*, 197, 1-7.
Illich, I. 1973. *Tools for conviviality*. Marion Boyars, London.
Hofkirchner, W. 2020b. Intelligence, Artificial Intelligence and Wisdom in the Global Sustainable Information Society, in: *proceedings*, 47, 39, 1-4, https://doi.org/10.3390/proceedings2020047039.
Lesch, H. The Capitalocene, 2018, https://www.youtube.com/watch?v=6wLlWWp8Vcg, accessed 20 August 2020.
Madrigal, A. C. 2011. Weekend Poem: All Watched Over by Machines of Loving Grace, in: *The Atlantic*, https://www.theatlantic.com/technology/archive/2011/09/weekend-poem-all-watched-over-by-machines-of-loving-grace/245251/, accessed 11 August 2020.
Nida-Rümelin, J., Weidenfeld, N. 2018. *Digitaler Humanismus – Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. Piper, München.
Nowak, M., Highfield, R. 2011. *Super co-operators: evolution, altruism and human behaviour or why we need each other to succeed*. Canongate, Edinburgh.
Precht, R. D. 2020. *Künstliche Intelligenz und der Sinn des Lebens*. Goldmann, München.
Simanowski, R. 2020. *Todesalgorithmus. Das Dilemma der künstlichen Intelligenz*. Passagen, Wien.
Tomasello, M. 2014. *A Natural History of Human Thinking*. Harvard University Press, Cambridge, Massachusetts.
Tomasello, M. 2016. *A Natural History of Human Morality*. Harvard University Press, Cambridge, Massachusetts.
Vienna Manifesto on Digital Humanism, https://dighum.ec.tuwien.ac.at.
Wu, K. 2020. The Impact of Intelligent Society on Human Essence and the New Evolution of Humans, in: *proceedings*, 47, 44, 1-4, doi:10.3390/proceedings2020047044.