Digital Humanism and the Limits of Artificial Intelligence

Julian Nida-Rümelin

Abstract This chapter is programmatic in style and content. It describes some patterns and one central argument of that, what I take as the view of digital humanism and which we exposed in our book (Nida-Rümelin and Weidenfeld 2018). The central argument regards the critique of strong and weak AI. This chapter does not discuss the logical and metaphysical aspects of digital humanism that I take to be part of the broader context of the theory of reason (Nida-Rümelin 2020, Chaps. VI and VII).

I

The expression "Artificial Intelligence" (AI) is multifaceted and is used with different meanings. In the broadest and least problematic sense, AI denotes everything from computer-controlled processes, the calculation of functions, the solution of differential equations, logistical optimization, and robot control to "self-learning" systems, translation software, etc. The most problematic and radical conception of AI says that there is no categorical difference between computer-controlled processes and human thought processes. This position is often referred to as "strong AI." "Weak AI" then merely is the thesis that all thought and decision processes could in principle be simulated by computers. In other words, the difference between strong and weak AI is the difference between identification and simulation. From this perspective, strong AI is a program of disillusionment: What appears to us to be a characteristically human property is nothing but that which can be realized as a computer program. Digital humanism takes the opposite side.

J. Nida-Rümelin (✉)
Ludwig Maximilians Universität Munich, Munich, Germany
e-mail: julian.nida-ruemelin@lrz.uni-muenchen.de

H. Werthner et al. (eds.), *Perspectives on Digital Humanism*,
https://doi.org/10.1007/978-3-030-86144-5_10

II

The analytic philosopher John Searle (1980) has devised a famous thought experiment. Searle asks us to imagine yourself being a monolingual English speaker "locked in a room and given a large batch of Chinese writing" plus "a second batch of Chinese script" and "a set of rules" in English "for correlating the second batch with the first batch." The rules "correlate one set of formal symbols with another set of formal symbols": "formal" (or "syntactic") meaning you "can identify the symbols entirely by their shapes." A third batch of Chinese symbols and more instructions in English enable you "to correlate elements of this third batch with elements of the first two batches" and instruct you, thereby, "to give back certain sorts of Chinese symbols with certain sorts of shapes in response." *Those giving you the symbols* "call the first batch 'a script'" [a data structure with natural language processing applications], "they call the second batch 'a story,' and they call the third batch 'questions'"; the symbols you give back "they call. .. 'answers to the questions'"; "the set of rules in English. .. they call 'the program'": *you yourself* know none of this. Nevertheless, you "get so good at following the instructions" that "from the point of view of someone outside the room," your responses are "absolutely indistinguishable from those of Chinese speakers." Just by looking at your answers, nobody can tell you don't speak a word of Chinese. Outside in front of the slot, there is a native speaker of Chinese, who, having formulated the story and the questions and having received the answers, concludes that somebody must to be present in the room who also speaks Chinese.

The crucial element missing here is apparent: It is the understanding of the Chinese language. Even if a system—in this case the Chinese Room—is functionally equivalent to somebody who understands Chinese, the system does not yet itself understand Chinese. Understanding and speaking Chinese requires various kinds of knowledge. A person who speaks Chinese refers with specific terms to the corresponding objects. With specific utterances, she pursues certain—corresponding—aims. On the basis of what she has heard (in Chinese), she forms certain expectations, etc. The Chinese Room has none of these characteristics. It does not have any intentions; it has no expectations that prove that it speaks and understands Chinese. In other words, the Chinese Room simulates an understanding of Chinese without itself possessing a command of the Chinese language.

Years later, Searle (1990) radicalized this argument in connecting it with philosophical realism (Nida-Rümelin 2018), that is, the thesis that there is a world that exists regardless of whether it is observed or not. Signs only have a meaning for us, the sign users and sign interpreters. We ascribe meaning to certain letters or symbols by communicating, by agreeing that these letters or symbols stand for something. They have no meaning without these conventions. It is misleading to conceive the computer as a character-processing, or syntactic, machine that follows certain logical or grammatical rules. The computer is comprised of various elements that can be described by physics, and the computational processes are a sequence of electrodynamic and electrostatic states. To these states, signs are then ascribed, to which we

attribute certain interpretations and rules. The physical processes in the computer have no syntax, they do not "know" any logical or grammatical rules, and they are not even strings of characters. The syntactical interpretation is observer-relative. As syntactic structures are observer-relative, the world is not a computer. This argument is radical, simple, and accurate. It rests on a realist philosophy and a mechanistic interpretation of computers. Computers are that which they are materially: objects that can be completely described and explained using the methods of physics. Syntax is not a part of physics; physics describes no signs, no grammatical rules, no logical conclusions, and no algorithms. The computer simulates thought processes without thinking itself. Mental properties cannot be defined by behavioral characteristics. The model of the algorithmic machine, of mechanism, is unsuitable as a paradigm both for the physical world and as a paradigm for human thinking.

A realist conception is far more plausible than a behaviorist conception regarding mental states (Block 1981). Pains characterize a specific type of feelings that are unpleasant and that we usually seek to avoid. At the dentist, we make an effort to suppress any movement so that we do not interfere with the treatment, but by no means does this mean that we have no pain. Even the imaginary super-Spartan, who does not flinch even under severe pain, can have pain. It is simply absurd to equate "having pain" with certain behavioral patterns.

III

It can be shown that logical and mathematical proofs to a large extent cannot be based on algorithms, as students of formal logics learn early on in their study. Already the calculi of first-order predicate logic do not allow for algorithmic proof writing. The fundamental reason for this phenomenon, that more complex logical systems than propositional logic are not algorithmic in this sense, is Kurt Gödel's incompleteness theorem (Gödel 1931), the probably most important theorem of formal logic and meta-mathematics. This theorem shows that insight and intelligence in general cannot be grasped adequately within a machine paradigm (Lucas 1961). One can interpret Gödels theorem as the proof that the human mind does not work like an algorithm. Possibly even consciousness in general is based on incompleteness as Roger Penrose (1989) argues, but I remain up to now agnostic about this question, being however convinced that neither the world nor human beings function like a machine.

If humans were to act just as deterministically as Turing machines (Turing 1950), then genuine innovation itself would not be imaginable. If it was in principle possible to foresee what we do and believe in the future, genuine innovations would not exist. Disruptive innovations in knowledge and technology require that future knowledge and technology is not part of old knowledge and technology. The assumption of an all-comprising determinism is incompatible with true innovation (Popper 1951, 1972). It is more plausible to assume that the thesis of weak AI, the

thesis that all human deliberation can be simulated by software systems, is wrong, than to assume that there is no genuine innovation.

IV

Digital humanism advocates the employment of digital technologies in order to improve human living conditions and preserve ecological systems, also out of concern for the vital interests of future generations. At the same time, however, it vehemently opposes a supposedly autarchic technological development of digital transformation. It opposes the self-depreciation of human competence in deciding and acting in the form of strong and weak AI; it opposes the subsumption of human judgment and agency under the paradigm of a machine that generates determined outputs from given inputs.

The utopia of digital humanism demands a consistent departure from the paradigm of the machine. Neither nature as a whole nor humans should be conceived of as machines. The world is not a clock, and humans are not *automata.* Machines can expand, even potentiate, the scope of human agency and creative power. They can be used for the good and to the detriment of the development of humanity, but they cannot replace the human responsibility of individual agents and the cultural and social responsibility of human societies. Paradoxically, the responsibility of individuals and groups is broadened by machine technology and digital technologies. The expanded possibilities of interaction enabled through digital technologies and the development of communicative and interactive networks rather present new challenges for the ethos of responsibility, which the rational human being cannot evade by delegating responsibility to autonomous systems, be they robots or self-learning software systems.

Digital humanism retains the human conditions of responsible practice. It does not commit a category mistake. It does not ascribe mental properties based on a simulation of human behavior. Rather, it sharpens the criteria of human responsibility in the face of the availability of digital technologies, calls for an expansion of the ascription of responsibility to communication and interaction mediated by digital technologies, and does not allow the actual agents (and that is us humans) to duck away and pass responsibility on to a supposed autonomy of digital machines. Digital humanism is directed at strengthening human responsibility, at realizing the potentials of digitalization that relieve the burden of unnecessary knowledge and calculations in order to give people the possibility to concentrate on what is essential and contribute to a more humane and just future for humanity.

References

Block, Ned (1981), Psychologism and Behaviorism, The Philosophical Review 90 (1): 5–43.
Gödel, Kurt (1931), Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, Monatshefte für Mathematik und Physik 38: 173–198.
Lucas, John R. (1961), On Minds, Machines and Gödel, *Philosophy* 36: 112–127.
Nida-Rümelin, Julian (2018), Unaufgeregter Realismus. *Eine philosophische Streitschrift*, Paderborn: mentis.
Nida-Rümelin, Julian (2020), Eine Theorie praktischer Vernunft, Berlin/Boston: De Gruyter.
Nida-Rümelin, Julian, Weidenfeld, Nathalie (2018), *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*, München: Piper (italian translation: Milano: Franco Angeli 2019; korean translation: Pusan National University Press 2020).
Penrose, Roger (1989), The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, Oxford University Press.
Popper, Karl (1951), Indeterminism in Quantum Physics and Classical Physics, British Journal of Philosophy of Science 1: 179–188.
Popper, Karl (1972), Objective Knowledge, Oxford University Press.
Searle, John (1980), "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3 (3): 417–457.
Searle, John (1990), "Is the Brain a Digital Computer?", Proceedings and Addresses of the American Philosophical Association, 64 (3): 21–37.
Turing, Alain (1950): Computing Machinery and Intelligence, Mind 59: 433–460.