The present vision of AI... or the HAL syndrome

Pierre-Jean Benghozi, Hugues Chevalier

Benghozi, P.-J. and Chevalier, H. (2019), "The present vision of AI… or the HAL syndrome", *Digital Policy, Regulation and Governance*, Vol. 21 No. 3, pp. 322-328. https://doi.org/10.1108/DPRG-12-2018-0079

The present vision of AI… or the HAL syndrome[1]

Pierre-Jean Benghozi and Hugues Chevalier

Over sixteen years after the symbolic deadline of *2001: A Space Odyssey*, the power of fiction is such that the debates that accompanied the release of the film in 1969 are still emblematic at a time when artificial intelligence (AI) and machine learning are at the heart of the digital economy

The second part of the film is structured around the power and autonomy of HAL[2] 9000. As the astronauts are hibernating, this super-robot computer, on board the spaceship *Discovery 1*, is in charge of managing for several years the mission to search for a mysterious black monolith, a symbol of absolute knowledge. HAL does not display any physical presence, its active presence takes the form of a big red eye and a voice, which makes it as puzzling as it is disturbing.

We find it worth looking back at this masterpiece of science fiction in a paper dealing with AI as we deem that the perception of computing ushered in then by the film is still up-to-date. To that extent, one can speak of an "HAL syndrome". It is a sign of the pathology of analysts and commentators when they are dealing now with the stakes and risks of AI, that they stress the omnipotence of technologies and overstate the expected performance, the "autonomisation" of machines and the problems of human control and management, the anthropomorphism of the way to handle usages and the relationships with new tools.

Looking back at this syndrome, so as to understand its nature, appears necessary as the perception of new uses, the capacity to appropriate the digital dimension, the very conception of applications, terminals and infrastructures are highly structured by a shared vision of technologies that percolate and spread within society.

The fatherhood of HAL: John Irwin Good, Stanley Kubrick, and the shadow of Alan Turing

AI and machine learning are new buzzwords that permeate all segments of the economy from public administration to heath, e-commerce, industry, and connected objects.[3] Nevertheless, it may be useful to remind that the notion and the hype around AI are far from being new. Indeed, they emerge since at least the 50ies under various guises, the film itself and A.C. Clarke's short story, "The sentinel", which provided the basis for the film are good examples.

[1] This study was originally published in French as "La vision actuelle de l'IA … ou le syndrôme de HAL", in Blanc et al (2018), translated/adapted by Jean Paul Simon.

[2] One should note that the name was coined through shifting the letters of the acronym of the then dominant firm for computing, IBM, the same company that – an irony of history – is coming back strongly in the field of AI with Watson. HAL (Heuristically programmed ALgorithmic computer) is a sentient computer (or artificial general intelligence). HAL is capable of speech, speech recognition, facial recognition, natural language processing, lip reading, art appreciation, interpreting emotional behaviours, automated reasoning, and playing chess. See Clarke (1972).

[3] Many authors now speak of "Iconomics" in this regard (cf. Volle, 2014). "Icononomy" is a neologism coined to refer to a society whose economy, institutions and lifestyles are based on the synergy of microelectronics, software and the Internet.

Building on competences and breakthroughs accumulated during the second world war, Alan Turing wrote, in 1950, a founding paper on the future of computing (Turing, 1950). Alan Turing, mathematician and statistician, is considered as one of the "fathers of computing", as a prophet of AI. He died in 1954, but his work continued under the guidance of John Irwin Good, another, mathematician and statistician who outlived Turing until 2009. During the war, Turing and Good met at the secret centre of Bletchey Park where the coding of the German machine Enigma was being deciphered. Building a scientific and friendly collaboration, they worked together during the last months of the war to set up the first computer, Colossus. Good knew, having been involved in the work of his friend Turing, his perspectives and he shared the same vision of the future of computing.

Kubrick turned to this well-known and famous specialist to advise him on computing when he decided to shoot *2001: A Space Odyssey*. Kubrick thought of opening the film with interviews of scientists, among them J.I. Good,[4] so as to legitimise the hypothesis of a supercomputer like HAL.[5] Beyond the mere anecdote, this episode is meaningful as it reveals that it was indeed one of the best computing specialists who sketched out, at that time, the shape, the functionalities and the capacities of HAL, projecting the potentialities he was considering, as an expert, for computing in the future. J.I Good (1965) wrote: *"Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that man need ever make,* ***provided that the machine is docile enough to tell us how to keep it under control***". Our emphasis in the last sentence offers a strong caveat that contradicts the very theoretical arguments that precede it.

Therefore, what is at stake with the HAL syndrome hovers around the vision of the limitless possibilities of AI and of its necessary control by human beings. It also explains why in the scenario, Bowman, one of the astronauts, eventually unplugs HAL when he realizes that the machine is trying to take control of the spaceship, to kill the hibernating astronauts so as to as fully achieve its "mission". Bowman escapes and manages to disconnect all the memory modules of HAL. Was it admissible for viewers and, beyond that, society at large for HAL to win over the humans? Certainly not. However, this very attempt to do so generates questions on the verge of uneasiness and, in some cases, terror.

Once again, the fiction highlights strongly the deep ambiguity of AI. AI is a cognitive prosthesis that became unavoidable to deal with the amount of information now available over the Internet, with the information and communication technologies, with the data gathered on content and services platforms, as well as in complex systems. In other words, it stands for a mandatory auxiliary so as to address several challenges. It makes it possible to rely on a global vision of the environments in which we live, to face the complexity of choice and decision making in a context of hyper-supply, to master skills and knowledge that are more and more diverse and of different kinds while leaning on knowledge incorporated in the "intelligence" of terminals and algorithms (for instance in personal assistants like Apple's Siri or Amazon's Alexa). In such situations, how can we control the machine and the quality of the decisions? As Carr (2010) pointed out, the question is whether this knowledge transfer toward the machines is making us stupid, as it lowers our skills to understand and memorise.

[4] Cf. Dyson (1986), quoted in https://fr.wikipedia.org/wiki/2001,_l%27Odyss%C3%A9e_de_l%27espace.
[5] The idea was dropped as the film was already very long.

Or, alternatively, it allows us to get rid of low value skills (memorisation through learning "by heart"), so as to focus on what is essential and to grow enhanced aptitudes (supervision, browsing and fast reading), know (where) to find), or simply being able to momentarily leave some tasks (during a lengthy job for instance). In all these cases, the issue is how to redefine the new skills and human knowledge to accompany and allow an efficient and reasonable appropriation of the machines… and of HAL without being left with the only option to unplug or disconnect the machine.

HAL syndrome or syndromes?

What we call the HAL syndrome is a pathology that affects two entities: HAL per se, but also the users/viewers with a split psyche. HAL suffers from a syndrome that can be described as megalomaniac psychosis: i.e. to consider oneself as the best and the most efficient (including over human beings) to achieve a given purpose. Worse, to be ready to do anything to achieve its mission (whatever it could be). To that extent, HAL, a human invention, is nothing but the translation of the frenzied vision, mentioned before in the quotation, of its creator J.I. Good.

On the other hand, viewers and users of technologies can also suffer from an imagined frenzy; two-pronged and contradictory. Some are afraid of this over-powerful computer, able to kill people in cold blood, and to oust them. Others may marvel at HAL's achievements, at the feats of computing and AI. As imaginary reactions, these bear no relationship to the reality of HAL.

The perception of HAL as well as AI is twisted because of the strong ambiguity of the content of the machine. The machine and the algorithms that enable it to operate do not create a being gifted with autonomy, decision-making capacity, and capacity of emotions as an anthropomorphic vision may suggest. They are first and foremost the result of a human conception and programming language. Their weaknesses, insufficiencies and risks are less the output of the machine than of the conception. As computing people bluntly put it: "garbage in, garbage out"!

HAL, all too human, unhuman...

The film clearly states that HAL was conceived by humans, which triggers some embarrassment about its actions/reactions. Does HAL obey a deadly programme incorporated by its creators? Possibly, as there is no such thing as an innocent programmer. The algorithms a programmer creates are marked by their ethics, culture, obsessions and aversions. Has HAL been programmed to take control of the ship at a given time, whatever the costs, including coldly killing the astronauts? Has HAL been programmed to achieve a specific mission without the astronauts being informed? Eventually, is HAL acting because it was programmed to, or out of its own autonomy? This is one of the ambiguities of the film as HAL appears to be gifted with a conscience and free will. Even displaying a sensitivity and emotions as when its memories are unplugged, it begs like a sentenced person.

Because of these ambiguities, HAL reveals to the viewer the terrible mirror of the most extreme human behaviour: a total lack of empathy, despising men, serial crime, the quest of absolute power, and the prevalence over any other consideration. Here we have a questioning of the artificial dimension of intelligence, as much as of the kind of capacity of the machine for self-learning.

...Or a plain machine?

However, once the hood of the machine is lifted, HAL appears "naked", that is a plain technical assemblage. During the sequence when Bowman is unplugging HAL, the astronaut disconnects each component and slice of memory, one after the other. One can then perceive and hear how the computer is gradually shutting down, losing its substance, weakening its voice to end up with an atrophy of its basic memory (its DOS?), then being just able to sing a simplistic song *Daisy* (the first item placed in its memory?)[6]. HAL reveals itself for what it is: a simple machine with a computing memory, a hard drive, and processors suited to using and reproducing algorithms. What the film shows of the constitution of HAL is just a standard computerized automat: stacking up layers of memories, much like our present computers.

Once again, HAL helps us to understand that AI is often just an umbrella term that covers different configurations and systems (one even has to introduce a distinction between strong AI which starts the programme with simple tasks and ends with convoluted assignments and weak AI which mainly aims to reproduce faithfully the result of a specific predetermined behaviour and not giving space to unexpected application), made out of a tinkering of heterogeneous applications and services relying as much on the raw computing power (Big Data), as on self-learning (cf. AlphaGo) and connectivity (cf. cybersecurity). As noted by Ezratty (2016), it is not a technology per se but rather a set of solutions combining technical bricks, components and sensors, meant and designed according to specific usages and applications: connected cars, sales recommendations, and aid to diagnosis…

This dimensions of the bricolage of AI is without doubt one of the answers to the need to master and monitor referred to earlier. Indeed, the technological diversity of the various configurations of components of AI, as well as the unavoidable limits of their interoperability, offers for sure, the best protection against the all-mightiness of the HAL syndrome.

However, today we are still quite far from the performance attributed to HAL by the scriptwriters. The creators of HAL designed a multitasking computer with a big red eye that echoes "big brother is watching you". Being in charge of the entire management of the spaceship requires outrageous capacities, beyond the reach of computing in 1969, as well as of 2001 and even 2018. The HAL syndrome leads us to relativize the omnipotence granted to technology and willingly circulated by both digital companies and transhumanist thinkers that advocate the use of science and technology – including IT – to enhance the human condition, in particular by increasing the physical and mental capacities of human beings. So, no designer of computers would have either the idea nor the capacity to set up a completely multitasking computer, because of the limits of technology. It would be far too complex to programme and create the relevant architecture…without mentioning the risk of multiple bugs and conflicting algorithms, always possible when programming, and driving repetitive dysfunctions and at the end of the day, the inoperability of the machine.

Whether one deals with process management, objects and parameters supervision, aid to diagnosis, gathering of data or simulation, in each case a specific computing architecture is required, as well as proper algorithms that are not replicable for other uses. For instance, the AlphaGO computer is programmed to play the game of Go but it remains unable to learn how to sing. From a practical viewpoint, and presumably still for quite some time, each automated computer performs a specific task; and it is important that it can be achieved on a regular

[6] https://kottke.org/06/04/hal-daisy-2001
https://www.youtube.com/watch?v=UGsfwhb4-bQ

basis with a total reliability. Quantum computers could, in theory, multiply by ten the present capacity of computers in terms of speed and relevancy, but they are still sketches. Furthermore, these future computers will require the knowledge to develop ultra-complex algorithms. Their development could rely on programming robots, however the conception and the architecture of these algorithms could only be initially generated by human minds, with of course all the liabilities attached.

HAL, one fiction.

The power to inspire coming from HAL holds to its being part of an identifiable genre, fiction, a privileged container for projecting phantasms and fantasies about unknown domains and inaccessible territories. As early as in the antiquity, Homer populated his *Odyssey* with fantastic creatures (Cyclops, Medusa, sirens…), with many supernatural creatures crossing the path of man. Naturally when choosing the title of his film, Kubrick was thinking about the work of Homer and on how to update Ulysses' adventures to interstellar space[7] thanks to modern computing technologies (it is worth noting that HAL, like the Cyclops, is one-eyed). Although deprived of any physical appearance, HAL is connected to this array of fantastic creatures from fiction.[8] Good and Kubrick constantly play on the ambivalence between the personality allocated to HAL and the effectiveness of the machine of the same name. That same ambiguity prevents viewers from seeing HAL as it really is and is to be found in numerous discourses and analysis about AI technologies.

On the one side, one tends to project a hideous conscience over HAL, but stemming from its inventors: despising human, serial crime on behalf of the achievement of the task allocated, the appearance of rationality leading to the implementation of an unshared power. These features allocated to the machine point to a madness, unfortunately human, and to the systematic application of rules and approaches of rational efficiency.[9] However, the ambiguity of the reality of HAL lies with the fact that this machine, this programmed automat, this advanced technical tool, remains a computing tool that the astronaut can switch off whenever he likes.

HAL displays madness under the guise of technical rationality and makes the viewer fluctuate between being terrified by the monstrosity of a computer trying to outsmart human beings, and being fascinated by the presumed capacity of the computing machine. The real chimera of the film is to transform a computing machine into an almighty figure of fiction. The way the HAL syndrome operates in society lies in this overevaluation of the potential of machines, and in the belief of a supposedly all powerful computer science.

Box 1. From manpower to brainpower

The digital economy allowed the emergence of a new being: the "brainpower", the output of human-computer symbiosis. The repetitive part of physical or mental work is being done by the computer while the human brings his perception, his intuition, and his capacity of initiative. Linked with human judgment, the power of computing gives an unheard-of efficiency to the "brainpower", however the success of this symbiosis requires a very thorough analysis of the practical conditions of its implementation.

Fifty years after the release of film, the HAL syndrome, as it continues to influence our minds, becomes the basis of the questioning, concerns and enthusiasms triggered by AI.

[7] Science fiction films host a lot of them, some of them terrifying: *ET*, Alien, Darth Vader, etc.
[8] One can think of Golem, or Dr Jekyll and Mr Hyde.
[9] Remembering what Hanna Arendt (1951, p. 416) wrote about the murdering madness of the Nazi regime.

Therefore, it calls for a future reflection over the need and modalities of the regulation of the current technological dynamics. First, AI emphasizes the various challenges of development, as it triggers at the same time, issues of deployment, of how to accompany the applications, and how to insert within existing practices. In spite of the fact that the majority of the works focuses on some specific applications (voice and image recognition, finance, tracking frauds, and recommendations and predictions of the behaviours of consumers), the implementation of AI in the economy will be accompanied, as noted by Volle (2015) by more profound structural transformations – like the coming of "brainpower" (see box 1) – similar to the ones that took place a century ago with the rationalisation of the workforce. Besides, AI calls for the setting up of a regulatory framework able to cope with the pace of innovation in that field: the spreading of automatization to all sectors and types of work, the modalities of the implementation of the necessary human assistance and supervision, the creation and appropriation of monitoring and supervision tools, the capacity to control the fairness and the processes used by algorithms, and the anticipation of systemic risks. The task is obviously vast and, for networking technologies, interconnected and belonging to complex systems…It may not be sufficient to unplug HAL.

References:

Arendt H. (1951) The Origins of Totalitarianism, Penguin Books, https://books.google.co.uk/books?id=zGE8DgAAQBAJ&pg=PT416&redir_esc=y#v=onepage&q&f=false .

Blanc (P), Volle (M), Chevalier (H), et al, (ed) (2018), *Elucider l'Intelligence Artificielle*, ed. Institut de l'Iconomie, Paris, available through Amazon.

Carr, N. (2010), *The Shallows: What the Internet Is Doing to Our Brains*, W.W. Norton & Company.

Clarke, A. C. (1972), *The Lost Worlds of 2001*, Signet.

Dyson, F. J. (1979). *Disturbing the Universe.* Harper & Row, New York.

Ezratty, O. (2016), *Les avancées de l'Intelligence Artificielle*, 2016, http://www.oezratty.net/wordpress/2016/avancees-intelligence-artificielle-ebook/.

Good, J.I., Alt, F.L., and Rubinoff, M. (1965), (eds), "Speculations Concerning the First Ultraintelligent Machine", *Advances in Computers*, Vol. 6, pp. 31-88.

Rochet, C. and Volle, M. (2015), L'intelligence iconomique, De Boeck.

Turing, A. (1950), "Computing Machinery and Intelligence", *Mind*, Vol. 59, No. 236, pp. 433-460.

Volle, M. (2014), *iconomie*, Xerfi et Economica, Paris, http://www.volle.com/travaux/iconomie.pdf.