Viagens da Saudade

Coordenação

Maria Celeste Natário
Paulo Borges
Luís Lóia

Organização

Cláudia Sousa
Nuno Ribeiro
Rodrigo Araújo

Porto

2019

FICHA TÉCNICA

Título: **Viagens da Saudade**

Coordenação: Maria Celeste Natário
Paulo Borges
Luís Lóia

Organização: Cláudia Sousa
Nuno Ribeiro
Rodrigo Araújo

Editor: Universidade do Porto. Faculdade de Letras

Ano de edição: 2019

ISBN: 978-989-8969-26-2

DOI: https://doi.org/10.21747/978-989-8969-26-2/viag

URL: https://ler.letras.up.pt/site/default.aspx?qry=id022id1671&sum=sim

F. Luig*

Does AI Trigger Humanism in Humans?

Abstract: In cinema it is possible to acknowledge a particular relationship Humankind acquired towards Artificial Intelligence over time. In the early years, and still on black & white flics, you already sense that there is some kind of fear for the unknown as well as for an eventual superior reasoning that machines could turn out to exhibit someday. Today, and due to exponential technological progress, we tend to overhype our feelings regarding artificial (super)intelligence development to Godlike capabilities and doubtful intentions toward our own existence. Whichever the outcome, it seems that AI has somehow triggered forgotten values that we used to treasure as being typical human. This talk and its subsequent debate intended to discuss just that. Furthermore, to enrich the debate, intelligence explosion and its eventual outcome scenarios are presented shedding some light to the controversies exposed in contemporary science fiction cinema.
Keywords: Artificial Intelligence, Intelligence Explosion, Technological Singularity, Mind-Uploading, Humanism.

Despoleta a IA humanismo nos humanos?

Resumo: É possível reconhecer no cinema, uma relação particular entre o ser humano e a Inteligência Artificial ao longo dos tempos. Nos primórdios, e ainda em filmes a preto e branco, já se sente uma espécie de receio pelo desconhecido bem como por uma eventual superioridade que as máquinas pudessem vir a exibir um dia. Hoje em dia, e por causa do progresso exponencial da tecnologia, tendemos a sobrevalorizar os nossos sentimentos relativamente ao desenvolvimento da super(inteligência) artificial a divinas capacidades e duvidosas intenções para com a nossa própria existência. Qualquer que seja o desfecho, a IA aparenta ter despoletado valores esquecidos que tínhamos como tipicamente humanos. Esta palestra e o debate daí resultante tenciona discutir precisamente isso. Mais, e para enriquecer o debate, a chamada explosão da inteligência e os eventuais cenários daí resultantes são aqui apresentados evidenciando assim as controvérsias já expostas no cinema contemporâneo.
Palavras-Chave: Inteligência Artificial, Explosão de Inteligência, Singularidade Tecnológica, *Mind-Uploading*, Humanismo

* F. Luig @ Universidade de Évora;
F. Luig @ Instituto Português do Sangue e da Transplantação;
F. Luig @ inLET (infinite Life Extension Technologies)

Introduction

A chronological series of sci-fi classic blockbusters was presented in this talk in order to establish a trend in the way modern civilization regards the progress of AI generally depicted in the form of humanoid robots. In Fritz Lang's Metropolis, *the Maschinenmensch*, cinema's first android displayed self-aware manipulation and as such qualifying as an intelligent being. This amazing representation of AI by the year 1927, which goes to say almost 100 years ago, truly stands for the beginning of an AI Era in pop-culture. Some of the iconic sci-fi movies are subsequently explored in the dialectics concerning humanism versus artificial intelligence.

The human-AI relationship in modern Sci-FI cinema

Before the eighties, AI characters didn't stand out as brilliant intelligent beings on the silver screen. Instead, one gets the impression they were merely efficient programming and calculating machines and took roles mostly as comic relief of some kind. Robby in Forbidden Planet (1956) and C3PO and R2D2 in Star Wars (1977), are clumsy robots clearly based on Metropolis in design and in function and do give the comedian style an extra laugh to the audience instead of an impression of an advanced intelligent system machine of any kind.

In Ridley Scott's cult classic Blade Runner (1982), Rachael is one of the most iconic examples of an android with self-awareness and can do anything a human does, even almost pass a Turing Test by tricking Harrison Ford's character Decker into believing she is for real. The other replicant, Roy, has the philosophical highlight when its lifespan of four years is almost up, desiring only what all living creatures want: not to die. Unlike the untrustworthy robots of Do Androids Dream of Electric Sheep? (Philip K. Dick) which supports the movie's script, Roy is the most human and alive character on screen including the ones playing real humans just like Decker only following orders in his slaughter of replicants. By this time it is fair to say that human society looks up to possible AI androids displaying humanlike features and values. It is almost as if humans would have had robotized themselves in routine day-to-day life and repetitive thoughts whereas the androids give rise to the meaning of life, deepful thoughts and attached values.

In the nineties, along with the uprise of the internet, grows the concept of an AI inside the web manipulating and eventually taming the human race. Movies such as the matrix ignite a whole new universe of possibilities in terms of *artificial superintelligence.* The AI creates a software The Matrix (1999) that imitates the real world, and uses it to replace the reality of humans. On the other

hand in the Terminator series, Skynet is a fictional neural net-based conscious group mind and artificial general intelligence (superintelligence) system that features as the main antagonist by unleashing a nuclear strike to terminate the human race as it entails humanity as a threat to its existence. The internet enhances the power of AI in extension and in reach, giving rise to a new kind of fearful entity that can spread its influence at the speed of light throughout the universe.

In the most recent years in movies such as Prometheus (2012), Ex-Machina (2015), AI has definitely been gaining momentum in terms of human behavior in all aspects inclusive in *emotional intelligence.* Is it just a side effect of the times we're living giving more importance to emotional intelligence or is it precisely because, again, AI apparently behaves the *sensitive way* humans should? The way we wanted humans to be? More human?

These androids (David / Ava / etc) exhibit evermore typical human features and attributes which ironically seem to lack in the real human counterpart characters. David is the kind of robot that wishes to be a human trying to act «human» by imitating Lawrence of Arabia which he watches over and over again while his human teammates are asleep during the trip. Ava, a femme fatale designed to manipulate men, includes the need for gender and sexuality in artificial intelligence. This AI is on a whole new level as it uses self awareness, imagination, manipulation, sexuality, empathy to dupe the human interrogator and its creator and involves using the Turing Test for proving capabilities and consciousness of an Artificial Intelligence.

Although dating back to 1968, 2001 - A Space Odyssey, truly stands out as the mark *per se*, which explains why it came out of the chronological order in this talk. HAL represents the perfect machine consciousness if it should ever develop. A computer that controls the mission to a singularity of unknown origin on a Jupiter moon. When two astronauts decide to take HAL offline, he reacts as any other living thing does by defending his right to exist. For Arthur C Clarke and Stanley Kubrick, HAL is the climax in humans evolutionary process of understanding the universe with the advent of controlling fire and shows technology has also evolved to the point where it dismisses the need for its creator. The consciousness exhibited by AIs such as HAL reflects human's fear but also the awareness that other intelligences could wake up and eventually display humanlike features too. Does it trigger humanism?

The movie AI (2000) also sets another benchmark. David, a robot-child designed to love, to do something so human only humans should be able to, is basically just a boy that wants to be human as much as Pinocchio is, to be a 'normal' boy, and that wants to be acknowledged in being so. As

if this acknowledgement would certify his humanism and this would then make him feel more real, more alive? Towards the end of the plot, in a somewhat distant future, humanity has meanwhile evolved into this digital/information like humanoid superintelligent being that ironically finds in David, a robot for that matter, the best and only representation of what humans used to be. To represent humankind through an android boy-look-alike when all humans are gone is at least bewildering. Does this not trigger humanism in humans?

Humanism is commonly defined as a belief that human needs and values are more important than religious beliefs, or the needs and desires of humans. Humanism is a philosophical and ethical stance that emphasizes the value and agency of human beings, individually and collectively, and generally prefers critical thinking and evidence (rationalism and empiricism) over acceptance of dogma or superstition (Cambridge English Dictionary). Humanism was important during the Renaissance because it was a time when people changed how they thought about humanity, art and philosophy. Humanism may become important again with the uprise of AI.

In all these movies (mainly the latter ones), the AI eventually becomes aware of its superior capabilities, meaning it surely developed some kind of consciousness by the time. It then usually turns against human race out of greed, arrogance or threat to existence, etc (all basic human instincts). This of course also reflects human's bias in anthropocentrism in assuming intelligent machines would or should behave like humans do. We can't expect them to do just what they were designed to. They will think, analyze and decide new targets in their own development and may or may not exhibit human behavior, and as such destructive, but with unimaginable power.

Intelligence Explosion and The Technological Singularity

(Technology grows exponentially and reaches infinity in finite time)

The knowledge doubling curve is accelerating exponentially and using Moore's Law we can extrapolate that in terms of raw processing power (petaflops), computer processing power will meet or exceed that of the human mind before 2020 which still does not mean we have a computer equal to the human mind. Software plays a key role in both processing power (MIPS) and AI (Delmonte, 2013). The distinction between artificial intelligence and artificial general intelligence separates thinking-machines programmed to be problem solving, task orienting mechanisms (narrow AI) from general-purpose systems with developing intelligence comparable to the human mind (AGI) (Goertzel, 2012). Once an AI system with roughly human-level general intelligence

is created, an intelligence explosion involving the relatively rapid creation of increasingly more generally intelligent AI systems will very likely ensue, resulting in the rapid emergence of dramatically superhuman intelligences (Goertzel, 2015).

Humankind is arguably approaching a technological singularity in the next few decades, based on the acceleration of scientific and technological progress, mainly through areas such as nanotechnology, biotechnology and artificial intelligence. The Event, an instant in the future linked to the development of a superintelligence (Bostrom, 2014), is generally tied to exponential growth in various technologies (with Moore's Law being a prominent example) as a basis for predicting that the singularity is likely to happen sometime within the 21st century (Kurzweil, 2005) and it generally entails three possible perspectives:

1. Accelerating Change (R. Kurzweil): Technological change feeds on itself and therefore accelerates. Change today is faster than it was 100 years ago. This trend advocates that in few decades computing power should exceed that of «unenhanced» human brains which allows to predict superhuman artificial intelligence. In a post-singularity world an intelligence explosion leading to a superintelligence would eventually transcend our brains beyond our bodies leaving at some time no distinction between human and machine. Acceleration relates quantitative measures of technological processes of evolution: milestones or paradigm shifts proving an accelerating pace of change as an upwards-curved plot leading to a discontinuity.

2. Intelligence Explosion (I.J.Good, Eliezer Yudkowsky): Intelligence makes technology and If technology improves intelligence then a positive feedback is triggered. Augmented intelligence will enhance itself further on in a loop like a chain reaction exponentially leading to superintelligence whether from an (AI) Artificial Intelligence development whether from an (IA) Intelligence Augmentation of the human brain.

3. Event Horizon (V.Vinge): Man has in the past been the reference in and of intelligence. Everything we have, we do because of our intelligence and technology will enhance it through brain-computer interfaces. This will create an uncertainty in the future that is unintelligible. An event horizon. The discontinuity accounts for a turning-point in human history as in Von Neumann's definition (some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.). The comes from the discontinuity of a gravitational singularity at the centre of black holes at which quantities become infinite or meaningless. Superintelligence stands for that level of intelligence in which traditional measures become

ineffective and its emergence marks an event horizon, because even «a tiny increment in problem-solving ability and group coordination is why we left the other apes in the dust» (Sandberg 2014): a discontinuity in our ontological and epistemological account of our existence.

In these sci-fi movies (also in those not covered in *the talk*) one can through the AI level achieved detect that a technological singularity of some kind must have occurred already, although by the time the movies were made (or even outlined) the term and concept itself had not been identified yet as the event described above and currently dubbed so. From here we can assume that, as growth of computing power and AI development has been taking form, human western culture has through literature, cinema and other art forms (music, etc) expressed its dreams but also its fears and somehow disclosed a growing feeling through time and progress. A feeling I called humanism, in the sense that it reveals human values, typical human behaviors and therefore of course also basic aggressive instincts. Existentialism and darwinian competitiveness. All of the above reflected in the future humanlike AIs characterized in cinema. As AIs evolve, in films as in reality, do they not trigger humanism by acting and being so human?

Mind Uploading and the Consciousness Conundrum

Most proposed methods for creating superhuman or transhuman minds fall into one of two categories: intelligence amplification of human brains [IA] and development of artificial (general) intelligence [AI]. The means speculated to produce intelligence augmentation are numerous, and include bioengineering, neurotropic drugs, AI assistants, brain-computer interfaces and mind-uploading. Furthermore, multiple paths to an intelligence explosion makes a singularity more likely to happen; for a singularity to not occur they would all have to fail. Its occurrence probability therefore increments with the variety of technological solutions developing. «The potential in achieving a superintelligence in a machine substrate is vastly greater than in a biological substrate though. Biological humans even if enhanced, will be outclassed» (Bostrom, 2012). In the artificial (super)intelligence scenario [AI], the emergence of human-level AI or even higher, forecasts the possibility of digital minds in the near future and relates to AI becoming one of the greatest potential threats to human existence (Sandberg, 2014) or «the worst thing to happen to humanity in history» (Hawking et al. 2014). This scenario poses a dystopian outcome for our society whereas the [IA] (intelligence amplification) resulting from the merger of our biology with technology (Kurzweil, 2005) would stand more for an utopian outcome with a human future overcoming

disease, aging, hunger etc… Either way, in both scenarios or possible outcomes from the human point of view, there would always be technical ground for mind-uploading (process by which the mind, a collection of memories, personality of a specific individual, is transferred from its original biological brain to an artificial computational substrate) and this would by itself stand for a variety of issues such as personal identity and the consciousness conundrum, as presented and discussed in this talk. Assuming uploading of the mind and everything it stands for possible in technical ways (by WBE for instance), what would have been transferred? Would it be an indistinguishable copy (numeric identity)? Is a copy the same self? Does it transfer a personal identity as the kind that preserves continuity of consciousness (continuation of subjective experience within the same entity through time)?

One of the main technological preconditions for mind-uploading techniques to take place depend highly on the feasibility of a Whole Brain Emulation (WBE). Exploring the Brain using computer technology to integrate everything and providing data about everything on a molecular level is a quite recent entrepreneurship in our history so it should not shock us that the Brain itself remains our greatest mystery and decoding it eventually our grand achievement. The HBP (Human Brain Project) entails a number of research studies at different levels of organization and development and aims to upload the entire mouse brain still this decade and upload parts of the human brain in the next 10 years. Progress in computing has vastly increased our ability to collect, store, analyze, and communicate information allowing us to do things our natural bodies and minds cannot. Exascale computers are expected to deliver cellular level simulations of the complete human brain with dynamic switching to molecular-level simulation of parts of the brain when required. This approach of working backwards from measurements of the functioning system to engineer models of how that system works is called reverse engineering.

Neuroscientists combined with computer scientists have been developing tools and algorithms, artificial multi-level brain-like neural networks similar to the ones measured in the brain itself. Reverse engineering will show how neurons connect rendering us a virtual interaction as is done in vivo and in vitro experiments. Software that could fully mimic human brains at various levels would most certainly lead to WBE whether or not these systems were to possess mental states as long as functional similarity would have been achieved. The current roadmap on the field suggests such emulations as viable by mid-century and in this manner boosting development of neuroprosthetic devices (Sandberg & Bostrom, 2008). Neuromorphic chips already in

development will take the Human Brain Project to a level of replicating cognitive capabilities (Markram, 2013). Sandberg has thoroughly discussed the feasibility of a WBE as well as the criteria by which emulations are validated. Only simulations achieving full functional equivalence, meaning all relevant properties of the original system being replicated (and observable), would in the long term represent a WBE which corresponds to a structural validity (Sandberg, 2013).

The Consciousness Conundrum

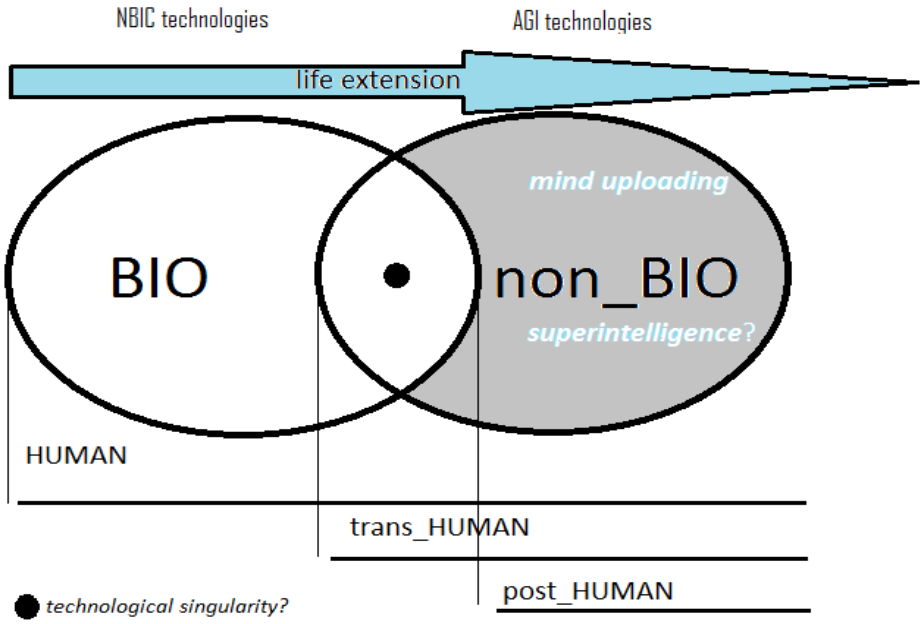
One of the major reasons of our human natural suspicion towards an upload comes from the fact that we put great value in consciousness. It does represent, to some extent, our main character of personality. It is our identity and one may even claim that without it we wouldn't be ourselves anymore. Yet in reality up to date there still is no true scientific explanation of what it even is. Daniel Dennett, as a notorious philosopher and cognitive scientist, dedicated more than 30 years to the subject and claims that there is no place in the brain where consciousness resides and that consciousness does not flow at all. There is no single 'stream' of consciousness and the continuity of consciousness is an illusion (Mental Darwinism, 1991 D.Dennett). In Häggströms thought experiment on the issue, humans base their judgement in other people being conscious or not on speech and behaviour rather than on their anatomy (Häggström 2016). In this manner we might be tricked into thinking an upload would be conscious by behaving as a conscious being, just like Alan Turing claimed more than 50 years ago when he considered machines could become intelligent. It should then be reasonable at least to assume that any emulated system could have the same material properties as the original system so we should treat it accordingly (Sandberg 2014). The CToM (Computational Theory Of Mind) states that what matters for consciousness is not the material substance itself but its organization and that the organization that produces consciousness is the right kind of information processing (Häggström, 2016).

But this doesn't help us in the personal identity issue since personal identity is not an organizational invariant. A clone, or even an identical twin, still would be a different person, numerically, even though qualitatively they could even be identical to the molecular level, we still would have two different identities. This would mean that even though an uploading could preserve an organizational state such as consciousness it still wouldn't mean it would preserve a personal identity. So it wouldn't also help to replace every single molecule. Even if fundamental physics

were to allow it, which doesn't seem to be the case, particles do not have or transport identities and the atoms we are made of keep changing over relatively small and definite periods of time and we keep on being ourselves anyhow. What persists is the pattern of organization not the individual molecules themselves. It is the cognitive structure of the brain (i.e. information) rather than the physical matter that counts; therefore any machine that duplicates the cognitive architecture of the brain will be conscious as the original brain (Cerullo 2015).

Conclusive Remarks

If consciousness is in fact an organizational invariant (systems with the same patterns of causal organization have the same states of consciousness) no matter whether that organization is implemented in neurons, in silicon or in some other substrate then it should be preserved in a computer simulation, thus rendering conscious states as the original system does (Chalmers, 2010).



the risk of the human being becoming *non-human* (or mainly not biological);
the dystopian risk of a divided society;
the existential risk.

All of the risks outlined in the schematics presented above are current issues on the social, cultural and political agendas. All of them are (and/or have been) explored in cinema. Depending on the feasibility of a whole brain emulation (WBE) or a neural network emulation, either by gradual

replacement of brain parts or by scanning, mind uploading would be achievable and substrate independent minds (SIMs) could come real. As humankind progresses in time towards a post-human Era, technologies seem to fuse efforts in transforming our main biological entity into an eventual non-biological substrate containing our essence. Our identity would in a digital form be preserved. Or maybe not. This exact dilemma directly relates to the consciousness conundrum which in turn opens a new debate around humanism as exposed and discussed in this paper.

Bibliographical references

BOSTROM, Nick. (2014), *Superintelligence, Paths, Dangers, Strategies*, Oxford: Oxford University Press.

CERULLO, Michael. (2015), *Uploading and Branching Identity*, Berlin: Springer.

CHALMERS, David. (2010), «The Singularity, A Philosophical Analysis», *Journal of Consciousness Studies*, Volume 17.

DELLMONTE, Louis. (2013). *The Artificial Intelligence Revolution (Will Artificial Intelligence serve us or replace us?*).

DENNETT, Daniel. (2018), *From Bacteria to Bach and Back: The Evolution of Minds*, London: Penguin.

GOERTZEL, Ben. (2012), *Why an Intelligence Explosion is Probable* (Singularity Hypothesis), Berlin: Springer Verlag.

HAGGSTROM, Olle (2016), *Aspects of Mind Uploading*, Chalmers University of Technology and the Institute for Future Studies, Sweden: Gothenburg.

KOENE, Randal. (2012), *Embracing competitive balance: The case for substrate-independent minds and whole brain emulation*, Singularity Hypotheses - A Scientific and Philosophical Assessment, Berlin: Springer Verlag.

KURZWEIL, RAY. (2005), *The Singularity is Near, When Humans Transcend Biology*, London: Penguin Books.

KURZWEIL, RAY. (2012), *How to Create a Mind, the secret of human thought revealed*, New York: Viking Penguin.

MARKRAM, H. (2011). *Introducing the Human Brain Project*, Elsevier.

SANDBERG, Anders. (2013). *Feasibility of whole brain emulation* in Vincent C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence*, Berlin: Springer (SAPERE).