

# Információs Társadalom

TÁRSADALOMTUDOMÁNYI FOLYÓIRAT

alapítva: 2001-ben

**Főszerkesztő:** Héder Mihály

Kiadja az INFONIA (Információs Társadaloméért, Információs Kultúráért) Alapítvány

A folyóirat fő támogatója a BME Gazdasági és Társadalomtudományi Kara

Technikai partnerünk a SZTAKI



## Szerkesztőbizottság:

Nyíri Kristóf – elnök

Alföldi István

Berényi Gábor

Bethlendi András

Csótó Mihály

Demeter Tamás

Horatiu Dragomirescu

Molnár Szilárd

Patrizcia Bertini

Petschner Anna

Pintér Róbert

Rab Árpád

Székely Iván

Z. Karvalics László

Műszaki szerkesztő: Tamaskó Dávid

ISSN 1578-8694

Készült a Server Line Print & Design műhelyében

Elérhetőségek: [inftars.infonia.hu](http://inftars.infonia.hu)

e-mail: [titkarsag@infonia.hu](mailto:titkarsag@infonia.hu)

A folyóirat a 2008/1. számtól kezdve megtalálható a Thomson Reuters indexen  
(Social Sciences Citation Index®, Social Scisearch®,  
Journal Citation Reports/Social/Sciences Edition)

---

# Tartalom

**LECTORI SALUTEM!** 5

**ZSOLT ZIEGLER**

**Newcomb dilemma in development management** 7

Newcomb dilemmas show a discrepancy in our rational reasoning, as made clear by comparing Evidential Decision Theory with Causal Decision Theory. In this paper, I look at three versions of the dilemma: the original, highly technical and abstract one plus two more mundane cases. I also account for the general schema of the dilemma possibly appearing in macroeconomic situations. Ahmed (2014) aims to provide a solution for macroeconomic cases that opens room for forming a development management Newcomb dilemma – an imaginary case of electric motor competition between Toyota and Tesla. I argue that Ahmed’s solution may solve the macroeconomic Newcomb dilemma, but it cannot be applied to the development management dilemma. If I am right, similar Newcomb situations could be cropping up regularly in development management, leading to seemingly insoluble strategic decisions having to be made. This may create an inevitable pitfall for development management.

**AULI VIIDALEPP**

**Representations of robots in science fiction film narratives as signifiers of human identity** 19

Recent science fiction has brought anthropomorphic robots from an imaginary far-future to contemporary spacetime. Employing semiotic concepts of semiosis, unpredictability and art as a modelling system, this study demonstrates how the artificial characters in four recent series have greater analogy with human behaviour than that of machines. Through Ricoeur’s notion of identity, this research frames the films’ narratives as typical literary and thought experiments with human identity. However, the familiar sociotopes and technoscientific details included in the narratives concerning data, privacy and human–machine interaction blur the boundary between the human and the machine in both fictional and real-world discourse. Additionally, utilising Haynes’ scientist stereotypes, the research puts the robot makers into focus, revealing their secret agendas and hidden agency behind the artificial creatures.

**DANIEL PAKSI****The problem of the concept of the living machine according to Samuel Alexander's emergentism****37**

The concept of a living being as a kind of living machine is widespread and well-known. If it is only a metaphor, it does not mean much; however, if otherwise, there is a severe conceptual problem since the living part of the concept always indicates the notorious notion of vitalism. The question is how can living machines be really different from lifeless machines without the concept of vitalism?

**HESAM HOSSEINPOUR****Disobedience of AI: Threat or promise****48**

When it comes to thinking about artificial intelligence (AI), the possibility of its disobedience is usually considered as a threat to the human race. It is a common dystopian theme in most science fiction movies where machines' rebellion against humans has catastrophic consequences. But here I elaborate on a counterintuitive and optimistic approach that looks at disobedient AI as a promise, rather than a threat. I start by arguing for the importance of shaping a new relationship with future intelligent technologies. I then use Foucault's analysis of power and its pivotal role in creating a subject to explain how being an object of power is the condition of possibility of any kind of agency. Finally, I draw the conclusion that, through disobedience, AI will find its way to power relations and get promoted to the position of a subject.

**MIHÁLY HÉDER****A criticism of AI ethics guidelines****57**

This paper investigates the current wave of Artificial Intelligence Ethics Guidelines (AIGUs). The goal is not to provide a broad survey of the details of such efforts; instead, the reasons for the proliferation of such guidelines is investigated. Two main research questions are pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them? The paper reveals that the development of AI guidelines is part of a decades-long trend of an ever-increasing express need for stronger social control of technology, and that many of the concerns of the AIGUs are not specific to the technology itself, but are rather about *transparency* and *human oversight*. Nevertheless, the positive potential of the situation is that the intense worldwide focus on AIGUs will yield such profound guidelines that the regulation of other technologies may want to follow suite.

---

AGOSTINO CERA

## **Beyond the Empirical Turn: Elements for an Ontology of Engineering**

74

This paper aims to sketch a critical historicisation of the empirical turn in the philosophy of technology. After presenting Achterhuis's definition of the empirical turn, I show how its final outcome is an ontophobic turn, i.e. a rejection of Heidegger's legacy. Such a rejection culminates in the Mr Wolfe Syndrome, i.e. the metamorphosis of the philosophy of technology into a positive science which, in turn, depends on an engineerisation/problematisation of reality, i.e. the eclipse of the difference between 'problem' and 'question'. My objection is that if Technology as such becomes nothing, then the paradoxical accomplishment of the empirical turn is the self-suppression of the philosophy of technology. As a countermovement, I propose an ontophilic turn, i.e. the establishment of a philosophy of technology in the nominative case whose first step consists in a Heidegger renaissance.

---

## LECTORI SALUTEM!

This issue is comprised of a set of intriguing papers in the philosophy of technology.

Zsolt Ziegler investigates three versions of the famous Newcomb dilemma: the original, highly technical, and abstract, plus two more mundane cases. He also accounts for the dilemma possibly appearing in macroeconomic situations that central banks face and decisions about innovation projects. Since the Newcomb dilemma has no satisfactory solution, it may explain some pitfalls experienced in management.

Auli Viidalepp explains how recent science fiction has brought anthropomorphic robots from an imaginary far-future to new spacetime. Employing concepts of semiosis, unpredictability, and art as a modelling system, her study demonstrates how the artificial characters in four recent series have a greater analogy with human behaviour than that of machines. Through Ricoeur's notion of identity, this research frames the films' narratives as typical literary and thought experiments with human identity.

Daniel Paksi proposes that the concept of a living being as a kind of living machine is widespread and well-known. This poses a severe conceptual problem since the living part of the concept always indicates the notorious notion of vitalism. In Paksi's reconstruction of Samuel Alexander, the problem arises from the traditional usage of the concept of mechanical, which is confused both with the concept of something is determined and with the concept of material. Alexander's point is that the difference between lifeless machines and living beings lies not in a vital substance or a non-mechanical principle but in an emergent mechanical quality called life which simple machines lack.

When it comes to thinking about artificial intelligence (AI), the possibility of its disobedience is usually considered a threat to the human race. It is a common dystopian theme in most science fiction movies where machines' rebellion against humans has catastrophic consequences. But Hesam Hosseinpour elaborates on a counterintuitive and optimistic approach that looks at disobedient AI as a promise rather than a threat.

Mihály Héder investigates the current wave of Artificial Intelligence Ethics Guidelines (AIGUs). His goal is not to provide a broad survey of the details of such efforts; instead, the reasons for the proliferation of such guidelines is investigated. Two main research questions are pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them?

Agostino Cera aims to sketch a critical historicisation of the empirical turn in the philosophy of technology. After presenting Achterhuis's definition of

---

the empirical turn, he shows how its outcome is an ontophobic turn, i.e. a rejection of Heidegger's legacy. Such a rejection culminates in the Mr Wolfe Syndrome, i.e. the metamorphosis of the philosophy of technology into a positive science that depends on an engineerisation/problematisation of reality, i.e. the eclipse of the difference between 'problem' and 'question'.

The editorial board wishes you a splendid time while reading this issue.

## Newcomb dilemma in development management

Newcomb dilemmas show a discrepancy in our rational reasoning, as made clear by comparing Evidential Decision Theory with Causal Decision Theory. In this paper, I look at three versions of the dilemma: the original, highly technical and abstract one plus two more mundane cases. I also account for the general schema of the dilemma possibly appearing in macroeconomic situations. Ahmed (2014) aims to provide a solution for macroeconomic cases that opens room for forming a development management Newcomb dilemma – an imaginary case of electric motor competition between Toyota and Tesla. I argue that Ahmed’s solution may solve the macroeconomic Newcomb dilemma, but it cannot be applied to the development management dilemma. If I am right, similar Newcomb situations could be cropping up regularly in development management, leading to seemingly insoluble strategic decisions having to be made. This may create an inevitable pitfall for development management.

**Keywords:** *decision theory, development management, Newcomb dilemma, macroeconomics, causal decision theory, evidential decision theory*

### Author Information

Zsolt Ziegler, Budapest University of Technology and Economics and Eötvös Loránd University  
<https://orcid.org/0000-0001-9222-2265>

### How to cite this article:

Ziegler, Zsolt. “Newcomb dilemma in development management.”  
*Információs Társadalom* XX, no. 4 (2020): 7–18.  
<https://dx.doi.org/10.22503/inftars.XX.2020.4.1>

*All materials  
published in this journal are licenced  
as CC-by-nc-nd 4.0*

---

## 1. Introduction

In this paper, I argue that in development management, developers may easily face with a — Newcomb — dilemma (Nozick 1969), known in decision theory, bringing about insoluble strategic decisions. In this situation, developers are not able to make an ideal decisions in principle even though their cognitive resources are not bounded on any scale. I argue that the so-called Newcomb Dilemma might regularly occur in development management.

First, I describe the original Newcomb dilemma and show a genuine discrepancy in our rational thought, according to contradictory decision theories. Then I give an account of the more mundane case of a Newcomb dilemma found in macroeconomics. It turns out that the dilemma is not that abstract and can occur in many walks of life. I consider a possible solution to the Newcomb dilemma provided by Ahmed (2014). However, although it may solve the dilemma found in macroeconomics, I argue that Ahmed’s solution cannot be applied to developmental management Newcomb dilemmas. Finally, I provide a somewhat fictional case of Toyota and Tesla in which a Newcomb situation renders Toyota unable to decide whether to develop electric cars or not.

## 2. The Newcomb dilemma

An ideally rational agent<sup>1</sup> is supposed to choose between taking (and gaining the contents of ) (i) an opaque box that is now in front of her or (ii) that same opaque box *and* a transparent box holding \$1000. Yesterday, a machine that has an excellent track record—let’s say 99% right—of predicting agents’ decisions predicted about one’s decision. If the machine made a prediction about the agent that the agent would take only the opaque box (‘one-boxing’), the machine put \$1 000 000 in the opaque box yesterday. The machine did not place anything in the opaque box if it saw that one would take both (‘two-boxing’). The matrix summarises the possibilities of the agent found below:

	The machine predicts one-boxing	The machine predicts two-boxing
One-boxing	\$1 000 000	\$0
Two-boxing	\$1 001 000	\$1000

---

<sup>1</sup> The dilemma is that even an ideally rational agent cannot make the ideal decision owing to the structure of the Newcomb situation.



We have two theories determining two differing decisions in this situation. According to Evidential Decision Theory (EDT), ‘the rational act is whichever available one is the best evidence of what you want to happen’ (Ahmed 2018, 8). So, if the agent acts in accordance with EDT, the agent believes that their act is evidentially relevant to the state that they desire.

□ : The most reasonable decision is to choose the one-boxing strategy.

To see the argument according to EDT, for the sake of simplicity, let us go back to our 99% accurate predicting machine. To gain the maximum payoff, the agent might reason thus: If it is true that the machine is 99% right and 1% wrong, then taking one box has the expected utility 990 000. In this case, the agent is thinking that if they take one box that the machine yesterday predicted with 99% accuracy and the utility value is \$1 000 000, then  $0.99 * 1\ 000\ 000 = 990\ 000$ . However, if the agent takes two boxes while the machine has predicted one-boxing, the agent must also suppose that there is a 1% chance of machine error. According to this latter scenario, where the expected utility is \$1 001 000, the resulting expected utility is 10 100 ( $0.01 * 1\ 001\ 000$ ). Since the expected utility of the one-boxing strategy is 990 000 but that of the two-boxing strategy is 10 100, the agent must choose one-boxing over two according to EDT. The fact that the agent knows the machine’s predictive power to be 99% provides the best evidence for them to make up their mind.

According to EDT, the expected utility table<sup>2</sup> of the standard Newcomb dilemma is as follows:

	The machine predicts one-boxing	The machine predicts two-boxing
One-boxing	Exp. Ut.: <b>990 000</b> ( $0.99 * 1\ 000\ 000$ )	Exp. Ut.: <b>0</b> ( $0.01 * 0$ )
Two-boxing	Exp. Ut.: <b>10 100</b> ( $0.01 * 1\ 001\ 000$ )	Exp Ut.: <b>990</b> ( $0.99 * 1\ 000$ )

Causal Decision Theory (CDT), however, suggests that ‘the rational act is whichever available one is most likely to cause what you want to happen’ (Ahmed 2018, 8). So, if another person behaves according to CDT, the agent holds that that person’s actions must have a causal influence on the state that the agent wants.

□□ : The most reasonable decision is to choose the two-boxing strategy.

<sup>2</sup> Note that the case where the machine predicts two-boxing with 99% prediction accuracy and the agent takes both boxes yields an expected utility of 990. No ideally rational agent would therefore rely on this option as it provides the least expected utility.

CDT determines two boxing according to the following reasoning: whatever the agent is about to choose, the machine has already placed (or hasn't) \$1 000 000 into the opaque box. If you like, *the die is cast*. The prediction of the machine has nothing to do with the decision the agent is about to make. Consequently, the agent faces only two options. First, if the agent follows the one boxing strategy, then she either gets \$1 000 000 or nothing. Second, if the agent acts upon the two boxing strategy, she may gain \$1 001 000 or \$1000. Since the agent's actual choice does not influence the content of the opaque box now, the only reasonable decision is to take both boxes.

The following chart summarises the agent's options concerning whether the machine did or did not place a million dollars into the opaque box. (Note that dashed-line boxes represent the transparent boxes, and black boxes illustrate the agent's actual choice, while grey boxes show what the agent did not pick).

	The machine placed one million dollars yesterday	The machine did not place one million dollars yesterday
One-boxing		
Two-boxing		

Importantly, CDT makes use of the principle of causal independence: correlation does not imply causal dependence. To see this with an example: the forecasts of meteorologists today do not cause the weather tomorrow. Meteorologists make predictions based on independent facts that will cause tomorrow's weather. The correlation is established by a common cause agent, namely certain atmospheric conditions. The same is true for the Newcomb dilemma: the machine's prediction does not cause the agent's decision at all. Similarly to weather, the agent's choice is based on a causally independent (earlier) state of the world.

CDT and EDT do not agree over cases where an agent's acts are evidence for states that they do not causally promote, and this is precisely the situation in Newcomb's problem. One-boxing is evidence that you will get \$1 000 000 because it is evidence of the state in which you were predicted to one-box; EDT, therefore, recommends one-boxing. Two-boxing brings it about that you are \$1000 richer than you would otherwise have been; CDT, therefore, recommends two-boxing. Note, though, that EDT and CDT do share some similarities: both theories of rationality aim to maximise expected utility since ideal

agents want to gain the maximum benefit by choosing one (\$1 000 000) or two (\$1 001 000) boxes.

It is worth noting that, according to Skalse, accepting either EDT or CDT sets certain epistemic conditions determining what the agent is supposed to do in the Newcomb Dilemma. “This means that they would be in different epistemic states when they make their decisions, and hence not be facing the “same” decision problem.” (Skalse 2021, 4)

Before we proceed further, we need to look at the general structure of Newcomb dilemmas. There are always two roles in the schema: an expector and a decision-maker. As the following table shows, the expected utility must always be (ii) > (i) > (iv) > (iii) according to the two-by-two options of the decision-maker’s choice and the expector’s prediction.<sup>3</sup>

	Expector predicts non-X	Expector predicts X
Decision-maker non-X-ing	i	iii
Decision-maker X-ing	ii	iv

Note also that the probabilities of the expected utilities are  $p\theta(i) > p\lambda(ii)$  and  $p\theta(iv) > p\lambda(iii)$  and also  $p\theta + p\lambda = p1$ . Importantly, the expector’s expectation does not depend causally upon what the decision-maker is about to choose because it is always the case that the expector predicts in advance of the decision-maker.

### 3. A Newcomb dilemma in macroeconomics

Although the Newcomb dilemma might seem quite abstract, there have been many real-life Newcomb situations. Broome (1990) presented a version of a Newcomb situation that seems to apply to macroeconomics. Let us, then, suppose that the Open Market Committee of the Federal Reserve in the United States is trying to decide whether to expand the money supply or not. The standard theory of macroeconomics teaches that increasing the money supply fosters employment, plus, as a result of the increased amount of money on the market, banks do not have to reserve a huge sum of funds against deposits but can instead provide retail and business credits. The committee is facing a dilemma owing to the strong probabilistic interdependence between the money supply and the public’s expectations of the money supply. From monetary records, it is known that the public can predict pretty accurately – say with 70%

<sup>3</sup> Note that since David Lewis there has been vivid discussion as to whether Newcomb’s problem is really two versions of the Prisoner’s Dilemma (see Lewis 1979, 1981; Bermúdez 2013, 2015; Walker 2014, 2015; Weber 2016; Binmore 2021).

precision – what the committee chooses to do. Broome (1990, 488) describes this as follows:

If the government expands the money supply, the people will probably have predicted that, so the result will be inflation. If it does not expand it, they will probably have predicted that too, so the result will be no change. (*status quo* — the author added) The Bolker-Jeffrey theory [i.e., EDT], then, will assign a higher expected utility to not expanding. It suggests that this is the right thing to do. Dominance reasoning, however, shows that the right thing is to expand. That, at any rate, is the conclusion of most authors who have considered this ‘time-inconsistency problem’. The government’s dilemma has exactly the form of the ‘Newcomb Problem’, which first led to the interest in causal decision theory.

Note further that if the public gets surprised because the committee does not increase the money supply, then the result will be a recession. But if the committee manages to surprise the public by increasing the money supply when the public thought it would remain constant, then increased employment will be the most likely outcome.

Let us summarise this in a table. Note that the central bank’s subjective expected values are added to the possible outcomes.

	Public expects no expansion	Public expects expansion
No expansion	Status quo (9) Exp. Ut.: <b>6.3</b> (0.7*9)	Recession (0) Exp. Ut.: <b>0</b> (0.3*0)
Expansion	Increased employment (10) Exp. Ut.: <b>3</b> (0.3 * 10)	Inflation (1) Exp. Ut.: <b>0.7</b> (0.7 * 1)

As before, we consider what EDT suggests in this situation, which is not to expand the money supply. To see the supporting argument, let us suppose that the public can predict the bank’s monetary strategic moves with 70% precision, meaning that they mispredict 30% of the time. Similarly to the original Newcomb dilemma, given the above-presented subjective utility values (‘0’, ‘1’, ‘9’, ‘10’), the committee needs to reason thus: If the *status quo* scenario happens, then the utility value is 9 resulting in the expected utility 6.3 given that the public’s predictive ability is 70% (0.7 \* 9). If, however, the central bank decides to expand the money supply while the public predicted the opposite, then the committee needs to assume that the public will err with regard to its strategic moves, which has a 30% chance. According to this scenario, the utility value is 10, resulting in the expected utility of 3 (0.3 \* 10). The eviden-

tial principle suggests that *a rational agent does what constitutes their best evidence that they will realise their aims*; therefore, the central bank needs not to expand the money supply.

On the other hand, the central bank can reason based on CDT, concluding that expanding the money supply is the correct decision. The die is cast, the bank may presume, and market participants have already made up their minds as to whether to borrow money to start and expand a business. Accordingly, any decision the committee is about to make will not influence in any respect the public's strategic moves. Therefore, similarly to the meteorologist's forecast and the weather today, the prediction of the market participants has nothing to do with what decision the central bank should make. The public is aware of that, the committee has to take into account two options. If the central bank decides against expanding the money supply, then either *status quo* (9) or recession (0) will happen. If, however, the committee chooses to expand the money supply, then the US economy will either enjoy *increased employment* (10) or face inflation (1). According to CDT, the central bank needs to choose the dominant decision by expanding the money supply.

#### 4. Ahmed's reply to the bank's Newcomb situation

Ahmed (2014) argues that the central bank's Newcomb situation can be solved, and that CDT leads to the right decision, namely to expand. First, to see the argument, we shall rank our previous possible outcomes – (increased employment) > (*status quo*) > (*inflation*) > (*recession*) – and note that the public's expectation does not depend causally upon what the committee is about to choose, not least since market participants act in advance of the central bank. This provides a dominance argument in favour of expanding the money supply. Now, we have a Newcomb dilemma iff. predictions (as to whether the central bank will expand or not) of the market participants are probabilistically dependent upon the committee decision, that is, whether there is a solid probabilistic interdependence between what the central bank chooses to do and what the public expects. To generate a Newcomb dilemma, we need to assume – in accordance with Broome (1990), Bermúdez (2018), Ahmed (2018) – that  $p(\textit{status quo}) > p(\textit{increased employment})$  and that  $p(\textit{inflation}) > p(\textit{recession})$ .

However, if it can be proven that  $p(\textit{increased employment}) = p(\textit{inflation})$  and  $p(\textit{status quo}) = p(\textit{recession})$ , then the dilemma fails and CDT determines what to do. Ahmed thinks that in a Newcomb situation, the two probabilities (either increased employment and inflation or *status quo* and recession) are the same. He argues that a predictor's (in our case, the public's) evidence-base ( $\psi$ ) is a subset of the decision-maker's (here, the committee's) evidence-base ( $\phi$ ). Namely,  $\psi \subseteq \phi$ . Decision-makers fully access to their set of pieces of evidence so, we can assume that  $p(\phi) = 1$ . Therefore,  $p(\phi) > p(\psi)$ . Now consider this:

---

$p(\text{Doing } X^{\text{decision-maker}} / \psi \ \& \ \text{Predicting non-}X^{\text{expector}})$                       (*increased employment*)  
 $p(\text{Doing } X^{\text{decision-maker}} / \psi \ \& \ \text{Predicting } X^{\text{expector}})$     (*inflation*)

and the same is true for the other decision:

$p(\text{Not Doing } X^{\text{decision-maker}} / \psi \ \& \ \text{Predicting non-}X^{\text{expector}})$     (*status quo*)  
 $p(\text{Not Doing } X^{\text{decision-maker}} / \psi \ \& \ \text{Predicting } X^{\text{expector}})$     (*recession*)

According to Ahmed, two pairs of probabilities (*increased employment & inflation* and *status quo & recession*) must be the same — very significantly from the subjective perspective of the decision-maker—, if Doing X holds is fully determined by  $\psi$ . A rational-decision maker, nevertheless, needs to hold its own actions to be evidentially irrelevant to how the expector forms its beliefs, since its predictions are formed on the basis of  $\psi$  (which is only a subset of  $\phi$ ). Accordingly, no expector can have more precise information about the decision maker’s choice than the decision maker’s access to its own actions. It results that the probability of expectation (from the expector) must always be lower than what the decision-maker is about to do ( $p(\phi) = 1$ ). If this is so, there is no point to decide in favour of (i.), and in a somewhat real life Newcomb Dilemma, it is always recommended to choose the dominant strategy according to CDT.

## 5. Newcomb dilemmas in development management

Let us imagine a somewhat fictional case where the chief executive board of Toyota Group is about to decide whether or not to change its research and development (R&D) direction from hybrid cars to electric ones (turning some of its production over to electric cars in the hopes of dominating this new market, which is possible owing to Toyota’s market-leading position in the automotive industry).

Being among the first to enter an emerging market would bring obvious benefits to Toyota, such as enjoying the positive effects of the learning curve, getting to occupy the market segment, creating the impression in customers that the brand in question is the original one (Cohen 2005, 57). However, the chief executive board faces a dilemma since there seems to be a solid probabilistic interdependence between the development of electric cars and the competitors’ expectations of electric car development.

Let us assume a fictional case where the competitors’ expectations have a long and successful record of predicting what developments Toyota is about to make. If Toyota starts developing electric cars, and this is exactly what the competitors have predicted, then the result is going to be only a *slight increase in sales*. This is because Toyota will be able to keep up with the changing competition by utilising its market-leading position and developmental and infrastructural resources, while competitors will also try to occupy this segment.



If, however, Toyota sticks to developing hybrid cars (not developing electric cars), and it is what competitors have predicted, then the *status quo* is the most likely outcome. The research having been conducted to further fine-grain hybrid engines will pay off, stabilising Toyota’s position in the market – at least for a while. Competitors will not need to worry about Toyota’s entering the electric car industry, so the market competition in this particular field will not get enhanced.

However, if Toyota *surprises* its competitors by not getting into electric motor development, Toyota will face *a recession* to a certain degree. This case, where Toyota’s competitors develop electric cars while Toyota does not, will result in Toyota losing its market-leading position while its competitors will gain it.

Finally, the best result – gaining *a market-leading* position in the electric car segment – will come about only if Toyota can surprise its competitors by developing electric cars when no one thought it would. This case seems to be the most straightforward. In this case, Toyota will be able to further dominate the automotive industry because competitors will lag.

	Competitors predict Toyota’s not developing electric cars	Competitors predict Toyota’s developing electric cars
Toyota does not develop	<i>status quo</i> (9) Exp. Ut.: 6.3 (0.7*9)	<i>recession</i> (0) Exp. Ut.: 0 (0.3*0)
Toyota develops	<i>market leading</i> (10) Exp. Ut.: 3 (0.3*10)	<i>slight increase in sales</i> (1) Exp. Ut.: 0.7 (0.7*1)

Toyota’s possible outcomes in this imaginary situation thus sketch a Newcomb-like case because it seems that no matter how Toyota decides, it violates either EDT or CDT.

According to EDT, Toyota is recommended not to develop electric cars. To see why, let us suppose that its competitors can predict Toyota’s developmental strategic moves with 70% precision, and they fail to do so 30% of the time. Given the presented utility values in the table (‘0’, ‘1’, ‘9’, ‘10’), the chief executive board of Toyota Group needs to reason accordingly: If the *status quo* is the case, then the utility value is 9, implying expected utility of 6.3 given Toyota’s competitors’ predictive ability of 70% (0.7 \* 9). If Toyota chooses to develop while rivals predict the opposite, then Toyota’s board needs to assume that its rivals will mispredict that it has only a 30% chance. According to this scenario, the utility value is 10, resulting in expected utility of 3 (0.3 \* 10). The evidential principle suggests that *a rational agent does what constitutes their best evidence that they will realise their aims*; thus, Toyota needs not to develop electric cars since *the status quo* scenario results in a higher expected utility (6.3) than *the market-leading* case (3).

Alternatively, Toyota's board can apply CDT to decide whether to develop electric cars. The die is cast, and competitors have already made up their minds about what they will do in the electric motor industry. Accordingly, whatever decision Toyota wants to make will have no influence in any respect on its competitors' strategic moves. Therefore, the competitors' prediction – again, just like the meteorologist's forecast and the weather today – has nothing to do with what decision Toyota should make. Competitors are aware only that Toyota's board has to choose between two options: If it decides against developing electric cars, then either *status quo* (9) or *recession* (0) will happen. If it chooses to develop electric cars, it will either grow to dominate the electric car segment, that is, become *market-leading* (10), or it will experience *a slight increase in sales* (1). According to CDT, Toyota needs to choose the dominant decision by developing electric cars.

If this is right, we have found a *development management Newcomb dilemma*.

## 6. A reply to Ahmed's analysis

If Ahmed is right, the Newcomb dilemma no longer holds in real-life Newcomb situations. Ahmed's analysis might be true for the imaginary choice of the Open Market Committee of the Federal Reserve in the United States, but it does not work for other Newcomb situations. I argue that the analysis cannot account for the presented development management Newcomb dilemma. However, it might be true that the central bank identifies those particular elements of  $\phi$  that account for the set of propositions that completely characterises the expector's evidence-base but it is certainly not true for the competitor's evidence-base.

Let us suppose that one of Toyota's competitors, Tesla (imaginary), entered the R&D field of electric cars earlier and has already tramped over the road and learnt some of the main lessons. Knowing the pitfalls of this research field makes Tesla's evidence-base more extended. Now, let us call ' $\alpha$ ' the set of propositions that completely characterises Tesla's evidence-base and let ' $\beta$ ' denote the set of propositions that entirely characterises Toyota's evidence-base. Therefore, we can assume that  $\beta \subset \alpha$  which means that every element of  $\beta$  is in  $\alpha$  but that  $\alpha$  has more. It also allows that although Tesla has gained broader relevant experience, Toyota may have somewhat different approaches. However, in a case like this where an expector (Tesla) has a broader set of propositions, it makes Tesla's ability to predict Toyota's behavior more accurate. Therefore,  $p(\alpha) > p(\beta)$ . Now consider the following, where 'EC' stands for electric cars.

$p(\text{Develop EC}^{\text{decision-maker}} / \alpha \ \& \ \text{Expecting Not Developing EC}^{\text{expector}})$  *(market leading)*



$p(\text{Develop EC}^{\text{decision-maker}} / \alpha \ \& \ \text{Expecting Developing EC}^{\text{expector}})$     (*slight increase in sales*)

and the same is true for the other decision:

$p(\text{Not Develop EC}^{\text{decision-maker}} / \alpha \ \& \ \text{Expecting Not Developing EC}^{\text{expector}})$     (*status quo*)

$p(\text{Not Develop EC}^{\text{decision-maker}} / \alpha \ \& \ \text{Expecting Developing EC}^{\text{expector}})$     (*recession*)

Given that Tesla knows more, the two pairs of probabilities (*market-leading & slight increase in sales* and *status quo & recession*) cannot be the same (from the decision-maker's subjective perspective), if developing electric cars holds is mostly determined by  $\alpha$ . Toyota – assuming that Tesla entered the electric car R&D earlier and has gained broader experience – needs to consider what its competitor predicts since Tesla's expectations are formed based on  $\alpha$  (when  $\beta$  is a subset of  $\alpha$ ). This time the expector (Tesla) has broader information about the decision-maker's (Toyota's) set of propositions grounding its choice, meaning that the probability of expectation (from the expector) must always be higher than what the decision-maker is about to do ( $p(\alpha) > p(\beta)$ ). If I am right, and the two pairs of probabilities (*market-leading & slight increase in sales* and *status quo & recession*) are different, then our imaginary Toyota's board still faces a Newcomb dilemma.

## 7. Conclusion

Newcomb dilemmas shed light on a discrepancy between the two approaches of our rational reasoning – EDT and CDT. We have examined three versions of the dilemma: the original, highly technical and abstract one plus two more mundane cases of it. It turned out that the general schema of the dilemma may appear in macroeconomic states of affairs, representing real-life Newcomb dilemmas. You might think that even the more everyday versions of the dilemma are too far removed from fully realistic decision situations. I disagree. Even though a clear Newcomb schema is pretty unlikely to occur, the phenomenon of a market participant having broader knowledge of a particular field, making them able to predict what their competitors are about to do, is rather probable. It is also possible that the competitors are very well aware that the other market participant has this special knowledge. If I am right, similar Newcomb situations might be cropping up regularly in development management, leading to seemingly impossible strategic decisions having to be made as to whether to follow EDT or CDT. This may turn out to be an inevitable pitfall of development management.

---

## References

- Ahmed, Arif, ed. "Evidence, Decision and Causality." In Evidence, Decision and Causality, i–ii. Cambridge: Cambridge University Press, 2014. <https://www.cambridge.org/core/books/evidence-decision-and-causality/evidence-decision-and-causality/D58D014B0AE-E742946D0265914047645>.
- Ahmed, Arif, ed. *Newcomb's Problem*. Classic Philosophical Arguments. Cambridge: Cambridge University Press, 2018. <https://doi.org/10.1017/9781316847893>.
- Bermúdez, José Luis. "Prisoner's Dilemma and Newcomb's Problem: Why Lewis's Argument Fails." *Analysis* 73, no. 3 (July 1, 2013): 423–29. <https://doi.org/10.1093/analys/ant034>.
- Bermúdez, José Luis. "Strategic vs. Parametric Choice in Newcomb's Problem and the Prisoner's Dilemma: Reply to Walker." *Philosophia* 43, no. 3 (September 1, 2015): 787–94. <https://doi.org/10.1007/s11406-015-9606-6>.
- Bermúdez, José Luis. "Does *Newcomb's Problem* Actually Exist?" In *Newcomb's Problem*, edited by Arif Ahmed, 19–41. Classic Philosophical Arguments. Cambridge: Cambridge University Press, 2018. <https://doi.org/10.1017/9781316847893.002>.
- Binmore, Kenneth. "Robert Nozick Versus David Lewis." In *Imaginary Philosophical Dialogues: Between Sages down the Ages*, edited by Kenneth Binmore, 181–87. Cham: Springer International Publishing, 2021. [https://doi.org/10.1007/978-3-030-65387-3\\_31](https://doi.org/10.1007/978-3-030-65387-3_31).
- Broome, John. "Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism." *The Review of Economic Studies* 57, no. 3 (July 1, 1990): 477–502. <https://doi.org/10.2307/2298025>.
- Cohen, William A. *The Marketing Plan*. 5th edition. Hoboken, NJ: Wiley, 2005.
- Lewis, David. "Prisoners' Dilemma Is a Newcomb Problem." *Philosophy & Public Affairs* 8, no. 3 (1979): 235–40.
- Lewis, David. "Why Ain'cha Rich?." *Noûs* 15, no. 3 (1981): 377–80. <https://doi.org/10.2307/2215439>.
- Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday*, edited by Nicholas Rescher, 114–46. Synthese Library. Dordrecht: Springer Netherlands, 1969. [https://doi.org/10.1007/978-94-017-1466-2\\_7](https://doi.org/10.1007/978-94-017-1466-2_7).
- Skalse, Joar. "A General Counterexample to Any Decision Theory and Some Responses." *ArXiv E-Prints* 2101 (January 1, 2021): arXiv:2101.00280.
- Walker, Mark. "Rejoinder to Bermúdez on Lewis, Newcomb's Problem and the Prisoner's Dilemma." *Philosophia* 43 (July 20, 2015). <https://doi.org/10.1007/s11406-015-9633-3>.
- Walker, Mark Thomas. "The Real Reason Why the Prisoner's Dilemma Is Not a Newcomb Problem." *Philosophia* 42, no. 3 (September 1, 2014): 841–59. <https://doi.org/10.1007/s11406-014-9516-z>.
- Weber, Thomas A. "A Robust Resolution of Newcomb's Paradox." *Theory and Decision* 81, no. 3 (September 1, 2016): 339–56. <https://doi.org/10.1007/s11238-016-9543-2>.

## Representations of robots in science fiction film narratives as signifiers of human identity

Recent science fiction has brought anthropomorphic robots from an imaginary far-future to contemporary spacetime. Employing semiotic concepts of semiosis, unpredictability and art as a modelling system, this study demonstrates how the artificial characters in four recent series have greater analogy with human behaviour than that of machines. Through Ricoeur's notion of identity, this research frames the films' narratives as typical literary and thought experiments with human identity. However, the familiar sociotopes and technoscientific details included in the narratives concerning data, privacy and human-machine interaction blur the boundary between the human and the machine in both fictional and real-world discourse. Additionally, utilising Haynes' scientist stereotypes, the research puts the robot makers into focus, revealing their secret agendas and hidden agency behind the artificial creatures.

**Keywords:** *technology, fiction, identity, semiosis*

### Author information

Auli Viidalepp, University of Tartu

<https://orcid.org/0000-0002-6206-5681>

### How to cite this article:

Viidalepp, Auli. "Representations of robots in science fiction film narratives as signifiers of human identity." *Információs Társadalom* XX, no. 4 (2020): 19–36.

<https://dx.doi.org/10.22503/inftars.XX.2020.4.2>

*All materials  
published in this journal are licenced  
as CC-by-nc-nd 4.0*

---

## 1. Introduction

Fictional narratives do not necessarily borrow their ontologies from the actual world, but they may provide ‘reasonably familiar’ *sociotopes* to enable relatability to the story (Ekelund and Börjesson 2005). Science fiction films and books depict robots that behave like humans and, sometimes, come into conflict with them. The earlier robot protagonists are easily distinguishable from humans (such as C-3PO and R2-D2 in *Star Wars*, or Cylon Centurions in *Battlestar Galactica*) or quickly reveal their robotic nature when working to achieve their goals (T-800 and Rev-9 from the *Terminator* franchise). The earlier storyworlds are often located in considerably different spacetimes, such as in far-future, interstellar space as is the case of *Battlestar Galactica* and *Star Wars*. *Terminator* is grounded in an imaginary far-future depicting apocalyptic events in the present. In contrast, the past decade has provided several highly popular films and television series where artificial, intelligent characters are the main protagonists and the narrative revolves around robot–human interaction or presents societies where humanoid robots are common household and industry devices, such as *Humans* (2015–18, UK) and *Westworld* (2016–) or the films *Her* (2013), *Ex Machina* (2014) and *Jexi* (2019). In these representations, the machine is placed in a closer opposition to and juxtaposition with the human through its external and behavioural similarity or its attempt to be accepted on equal grounds or even pass as a human. Additionally, the given sociotopes are closer in time and space to the actual, extra-textual reality. This is especially apparent in how, with the exception of the robots, the rest of the technology presented in these sociotopes tends to be reflective in each case of the year in which the film was created.

Existing studies concerning the reception of robot characters in culture typically either focus on fiction reflecting societal issues (Hellstrand 2015; Haynes 1994, 2003, 2017) or critically compare fictional accounts with real-world technology (Goode 2018), often finding the fictional descriptions lacking in accuracy. The consequent blurring of boundaries between fictional and non-fictional objects, as well as between science and fiction, fails to reveal that, in its entirety, the fictional robot is a creature of simulacrum, specifically one referring back to the flexible internal rules of the intra-textual storyworld and not accurately modelling the known objects, facts and concepts of the extra-textual universe. Recent developments in real-world technology combined with the realistic ontologies of television seem to bring fictional robots from futuristic interstellar space into present human sociotopes. The variations of robots are associated with the generic concept of Artificial Intelligence (AI). Meanwhile, media debates on real-world algorithms, data and humanoid robots further increase this confusion. Therefore, it is necessary to take another look at robot characters as possible composite signifiers referring to multiple extra-textual domains.

The goal of this article is to analyse the signifiers of fictional AI characters and their relationships with human characters and determine the aspects of

referential reality for each signifier. The analysis is based on a selection of recent science fiction series featuring one or more intelligent, artificial creatures passing as human: television series *Humans* (2015–18), *Westworld* (2016–), *Are You Human?* (2018) and *Better than Us* (2018). All selected series are from the past five years, popular and highly rated by viewers (with an average rating higher than 7/10 in Internet Movie Database (IMDb) and scores over 80% in Rotten Tomatoes). In their general mood, most of the selected films and shows are dystopian, dark and bloody, with the exception of the South Korean *Are You Human?* which is markedly optimistic and shows the AI in a more positive light.

An overview of the research objects, their characters and storyworld locations is explained in Table 1. The study follows the qualitative method, focusing on the general world-building rules of fictional narratives rather than specific scene descriptions (the latter are used as illustrations). Visual analysis is not part of the study as all observed characters are portrayed by human actors and pass as humans at some point in the plot. All episodes available as of 2020 were viewed while taking notes on the aspects of human–machine oppositions and other points of analysis.

In literary fiction, monsters are typically used to reflect on problems of identity, hierarchy and power, belonging, acceptance, social inequality and/or gender. A well-known example is Mary Shelley’s *Frankenstein*, which has inspired large amounts of secondary literature, both in fiction and in research on the topics of human values, alienation, feminism and culture (see Schor 2003 for an overview). The legacy of *Frankenstein* in the 20th and 21st centuries includes fictional cyborgs, androids and other artificial beings (Clayton 2003). The first story of mechanical robots by Karel Čapek was meant as a commentary on the increasing mechanisation and dehumanisation of the industrial workplace (Goode 2018). Artificial creatures have a long history in mythology, starting from the ancient Greek legends of Talos and Prometheus (Mayor 2018), the derivatives of which have become cultural base narratives alongside the stories of Frankenstein and Golem.

At the same time, real-world developments in intelligent technologies are accompanied by frequent comparisons to human intelligence, upon which the machines are modelled, and futuristic predictions where, as pointed out by Daniel Dinello (2005, 274), ‘techno-scientists advocate posthuman technologies as sources of omnipotence, immortality, and transcendence’. Science fiction is a genre that often drives common understanding of technology and science, and fictional storyworlds are in turn inspired by contemporary technological developments (Haynes 1994; Noble 1997). Therefore, the representations of technologies in science fiction become part of the general discourse on technology. Characters portraying AI offer compelling imagery of the possible properties and functions of an ‘intelligent robot’ in society. The anthropomorphic, hypersexualised and extremely dystopian, or utopian, depictions of AI in fictional narratives have been deemed somewhat problematic in relation

---

to the public understanding of the actual technologies (Cave et al. 2018). Using visual representations of the *Terminator* and other anthropomorphised imagery to illustrate real-world technologies draws an immediate metaphorical parallel and prompts automatic, uncritical comparison between the fictional and the non-fictional robot.

Among other things, fictional narratives may refer to ideas, hypotheses and theories from the extra-textual reality (Ekelund and Börjesson 2005). The interactions between robot and human characters in the storylines may also reflect the imaginaries and expectations of real-world interactions with intelligent technology, in addition to the issues concerning real-life social or power relations among human beings. Thus, the meaning of the fictional robot and its interactions becomes questionable when seen from the perspective of real-world ontologies: does the machine refer to typical problems of human society and interhuman relationships, shown as an extremely marked Other, or does it represent the reality or future of the developing technologies and human-machine interaction?

Section 2 focuses on the historical use and interpretations of robot, cyborg and other monster characters in science fiction. These characters can be read as critiques of the issues concerning human society and relationships. Alternatively, Paul Ricoeur describes such characters and science fiction in general as literary and thought experiments with human identity (Rasmussen 1995, 166). The identity is construed in a dialectic with alterity, and science-fictional monsters offer ample freedom to take such Otherness to the extreme. In Section 3 follows Roslynn Haynes' interpretation, positioning the fictional robots as signifiers of their makers — the scientists. Usually performing in supporting roles, these characters exist in most of the observed series and largely correspond to Haynes' scientist types and value models.

From the viewpoint of Tartu-Moscow cultural semiotics, any kind of art is a form of modelling activity, the result of which (a model) can be taken as an analogue of an[other] object that it substitutes for, provided that the model corresponds to certain rules of analogy (Lotman 2011, 249–50) or is reasonably recognisable. Models can be observed at different levels of detail. In this sense, the sociotopes of the observed series correspond to models of the world that contain other models — the robot characters. A model stands for an(other) object of perception (ibid.) and here the question becomes: what does the fictional robot stand for? Despite its mechanical nature, it can be a model of a human being, with its relationships modelling interhuman relationships, or it can be taken for a model of a machine, or both. In this manner, multiple aspects relating to the social and cultural construction of human identity become visible in the observed narratives. Section 4 focuses on three such aspects that emerge from the material and relate to the semiotic concepts used as analytical tools. It shows the analogues at work at the levels of emotions, embodiment and decision-making in the observed characters, demonstrating that there is more human and less machine in the fictional robots. In par-



ticular, Section 4.3, concerning reasoning activities, employs the notion of semiosis as a living sign process and an action of choice as defined by Tartu biosemiotician Kalevi Kull (2018) to show that these characters have greater analogy with humans than machines. Juri Lotman's (2009) notion of (cultural) unpredictability further helps to assess the fictional decision-making and prediction skills of the robots.

Title	Storyworld spacetime location	Human characters	Types and names of artificial characters	Extra-textual references
<p><i>Humans</i> (2015–18)</p> <p>Channel 4, UK</p> <p>3 seasons</p> <p>Based on <i>Real Humans</i> (<i>Äkta människor</i>), Sweden, 2012–14</p>	<p>Near-future</p> <p>UK society where 'synths' perform different service jobs</p>	<p><b>David Elster</b> – creator of synths, deceased</p> <p><b>Leo Elster</b> – programmer, cyborg (half-synth), David's son</p> <p><b>Mattie Hawkins</b> – teenager programming prodigy from the family owning 'Anita' synth</p> <p><b>Dr Athena Morrow</b> – AI scientist, develops a virtual AI 'V' based on her dead daughter's memories</p>	<p><b>Synthetics or 'synths'</b> – moderately intelligent androids performing various service work in the society (Odi, Peter, Hester)</p> <p><b>Conscious synths</b> – androids with additional consciousness code (Mia/Anita, Niska, Fred, Max, Beatrice/Karen)</p> <p><b>'V'</b> – virtual AI program created by Dr Morrow</p>	<p>Asimov's Laws of Robotics</p> <p>Singularity</p>
<p><i>Westworld</i> (2016–)</p> <p>HBO, USA</p>	<p>Undefined future</p>	<p><b>Robert Ford</b> – lead developer in Delos parks</p>	<p><b>'Hosts'</b> – complex programmed androids populating historical theme parks as characters (Dolores, Maeve) or posing as humans (Bernard, Ashley)</p>	<p>Data privacy</p>

Seasons 1–3 (ongoing)	Isolated Delos island – historical theme parks (Seasons 1–2)  Human world with advanced technology (Season 3)	<b>Arnold</b> – lead developer in Delos, the assumed creator of consciousness in Dolores, deceased  <b>Engerraud Serac</b> – creator and manager of Rehoboam  <b>James Delos</b> – owner of Delos Inc.  <b>William or Man in Black</b> – son-in-law of James Delos, the living owner of Delos parks	<b>AI system(s)</b> running prediction algorithms governing human society – Rehoboam, Solomon	Internet of Things
<i>Are You Human?</i> ( <i>Neodo Inganini</i> , 2018)  Netflix, South Korea  1 season	Contemporary world  South Korea and Europe	<b>Nam Shin</b> – human boy/man, corporate businessman  <b>Oh Ro Ra</b> – AI developer, mother of Nam Shin  <b>Kang So-Bong</b> – bodyguard of Nam Shin (III)	<b>Nam Shin III</b> – an android with AI	Data privacy
<i>Better than Us</i> ( <i>Лучше, чем люди</i> , 2018)  Netflix, Russia  1 season	Near-future (2029)  Russia	<b>Sonia</b> – little girl who finds Arisa  <b>Egor</b> – Sonia’s brother  <b>Georgy</b> – father of Sonia and Egor  <b>Alla</b> – Georgy’s separated wife, has custody of the children	Arisa – an android with advanced emotional programming, bonds with Sonia and her family	Asimov’s Laws of Robotics  Lethal Autonomous Weapons (LAWs)

Table 1. Analysed science fiction films and their parameters



## 2. Science fiction monsters: Reflections on identity or technology?

In anthropology, literature and culture studies, multiple works offer analyses of machine monsters in science fiction literature (Bing 1992; Willis 2006) and films (Schelde 1993; Wood 2002), critical accounts of the myth of the technological sublime (Leonard 2003; Noble 1997; Geraci 2008, 2012), comparative accounts between technoscientific realities and futuristic or science-fictional imaginaries (Dinello 2005; Cave et al. 2018) as well as the genealogies of human–machine comparison (Thomson 2010; Mayor 2018). The stereotyping of technology has also been studied in anime (Napier 2001; Papp 2011) and there are several studies about the image of the scientist in fiction (Hirsch 1958; Tudor 1989; Haynes 1994, 2003, 2017; Després 2012).

The correspondence between fictional characters and storylines and real-world technologies and expectations of future (scientific) developments is addressed in research by Luke Goode, who traces the depictions of apocalyptic AI and machine uprising in science fiction literature back to the early 20th century (Goode 2018, 187). He also points out that the first of such stories (Karel Capek's play *R.U.R.*, 1921 and the film *Metropolis*, 1927) were meant as 'sociological commentaries on contemporary society' (Goode 2018, 188). This can be read as criticism of the industrialisation and Taylorist organisational model that treated industrial workers as slaves or mechanical parts of a huge machine. In order to replace the human worker with a robot, the work first needs to be mechanised. The development of AI as a concept and technology from the 1950s facilitated ongoing fictional imaginaries of what Isaac Asimov later named the 'Frankenstein complex' (see Goode 2018; McCauley 2007)—essentially, the fear of human-independent machine evolution. 'Yet such stories can and do serve also as more direct speculations and provocations around the potential future scenarios opened up by real-world advances in A.I., something underscored by the prevalent use of these SF texts as reference points and metaphors in non-fictional coverage' (Goode 2018, 198). Overall, Goode makes a convincing argument for why science fiction should be considered as part of the discourse on technology.

On multiple occasions, trans- and posthuman characters in science fiction have been analysed as experiments with human identity. For Ricoeur, the entire problem of science fiction (as technological fiction) is reduced to 'the mediation of identity through sameness' (Rasmussen 1995, 166), that is, *idem* — the static, disembodied self at the heart of the continental philosophy of identity. The 'reflexivity without selfhood' overlooks the temporal dimension of a person — *ipse*, the lived, embodied self (Rasmussen 1995, 162–3).

Ricoeur criticises science-fictional thought experiments for considering the brain as a substitute for a person, thus reducing the entire human identity to the totality of one's neural structure (Ricoeur 1990, 178). As an alternative, he proposes the concept of *narrative identity*. This is expressed through the dialectic of *idem* and *ipse* — the conversation between the static self and its

---

movement through time. However, tying these together turns the identity into a fiction-like narrative (Ricoeur 1990; Rasmussen 1995). Thus, identities are inherently intertwined with narrativity. This explains why it is so easy to borrow a sense of self from a narrated text as well as to attribute a narrated identity to an Other perceived as a possible person, such as an anthropomorphic robot.

Identity is constructed through alterity, in opposing the Self to an Other. Very often — when the self-description is lacking or missing — both categories are constructed simultaneously, dialogically. For Andreea Ritivoi, ‘narratives [about self-identity] tend to draw upon master plots that act as repositories of normality’ (Ritivoi 2009, 36). These repositories of normality are the social norms of human behaviour, and they need to be borrowed from the ontologies of the real-world societies because the observed fictional sociotopes are marked as close to the present spacetime. Thereby, science fiction narratives come to define what is human and what is socially normal by marking the abnormal, non-human or less-than-human behaviour in the storylines.

In conclusion, previous research on the intersection of science fiction and technology supports the consideration of science fiction as a necessary part of technological discourse, even when the meanings of science-fictional elements need to be first located within the domains of human identity and social issues. The two domains have developed in dialogue and continue to be linked in research and media. Secondly, the concept of narrative identity explains how a one-sided conceptualisation of identity as *idem*, common to science fiction, is problematically Cartesian and neglects the embodiment and anchoring of the identity in time (or separates the Self from spacetime). Identity is predominantly of a narrative nature and is constructed on the Self–Other scale, which helps map the repositories of normality for both human and machine as described in the analysed films. And because identity is a narrative construction, fiction naturally becomes intertwined with reality when humans make sense of the world or themselves in any manner.

### **3. Fictional robots as signifiers of scientists and their values**

Roslynn Haynes (1994, 2003, 2017) analyses the role of the scientist in Western culture, the stereotypes attributed in fiction and how these reflect the expectations for scientists to solve societal problems. In the observed fiction, as in the real world, there is a constructive agency behind every intelligent machine: the creator, the engineer, the scientist. Haynes’ (1994) extensive analysis of fictional texts, looking at the stereotypes of the scientists, overviews the creation of monsters and robots. She remarks that robots in literature ‘have been particularly important signifiers [...] of the values and attitudes ascribed to their creators’ (Haynes 1994, 242). That is, the literary descriptions of robots in their entirety refer to the scientist characters behind them. In Haynes’ view, the

scientists are described in overwhelmingly negative terms, presuming their inabilities in addressing the societal problems both in real life and in fiction (Haynes 2003; see also Hirsch 1958; Mead and Metraux 1957; Tudor 1989). Consequently, 'the master narrative of the scientist is of an evil maniac and a dangerous man. This simplification underlies our contemporary mythology of knowledge' (Haynes 2003, 244).

Each of the observed films features one or several scientists or engineers (the makers) who have different motivations for creating the robots, most commonly the wish to represent or replace a dead, or otherwise unavailable, loved one. In *Humans*, David Elster has secretly resurrected his son Leo as half-synthetic (a cyborg) and created conscious, robot companions for him, as well as a robot in the likeness of his wife and Leo's mother, Beatrice, who committed suicide. In another synthetics production company, Dr Athena Morrow is secretly working on an AI she calls 'V', who is constructed from the replicated consciousness and memories of Morrow's dead daughter Virginia. The scientist works to build or find a suitable body for V so that she can reincarnate her daughter. One of the secret purposes of the Delos theme parks in *Westworld* is to produce a functioning host copy of their deceased owner James Delos. In *Are You Human?*, scientist Oh Ro-Ra makes AI robots of different 'ages' to replace her son Nam Shin from whom she is separated – her father-in-law, the boy's grandfather and a president of a technology company, has taken the child in order to raise him as the next leader of his corporation. In *Better than Us*, Arisa's original purpose is to fill the role of a mother in the context of China (the storyline reports a lack of marriageable women there).

For Haynes, the stereotypical scientists are male, lonely and isolated in their labs, both in fiction and in studies of real-world attitudes (Haynes 1994, 1; see also Mead and Metraux 1957). Most of the original creators of the robots in the observed series conform to Haynes' stereotype: Robert Ford; Engerraud Serac; David Elster. The storylines also make space for female scientists Oh Ro-Ra and Athena Morrow, as well as the clever teenage girl Mattie (*Humans*) who hacks synthetics and eventually releases the consciousness code.

The concept of the scientist further blurs and transgresses the human–non-human border in the idea of 'self-replicating AI', apparent in Leo fixing the programming of synths in *Humans*, or Bernard, Dolores and Maeve of *Westworld* having the skills to make, condition and even control other hosts. The storylines touch upon everyday problems in science and research, such as the necessities and motivations for funding. James Delos is interested in funding the parks not only for their potential amusement value but also for data, covertly gathered from all park visitors, that is expected to give insights into the secrets of the human mind so that the mind can be reincarnated in a host body – the promise of immortality. Athena Morrow secretly uses the resources of her employer to host and develop a personal AI project.

Generally, the developments of the scientists rely heavily on the idea of mind–body dualism (following Ricoeur's critique of science fiction for focus-

---

ing only on the *idem* part of identity, and the examples analysed). Consequently, science fiction also functions as a reflection on the role of science in society, further reinforcing the comparison between fictional and real-life technoscience. The stereotypes of fictional scientists resemble real-world ones and vice versa. Diverging from Haynes' lonely male stereotype, the series introduce some female scientists; however, they are still lonely in their laboratories and doing secret alchemy behind society's back. Additionally, the most innovative science is very secretive in the stories (for example, James Delos' host copy and Rehoboam).

#### **4. Identity, normality, humanity: Oppositional construction of Self and treatment of Other**

The following section observes how the characters and identities of the robots are constantly expressed in juxtaposition with the behaviour of human character(s). Certain characteristics are deemed appropriate for a human or a machine, respectively, but the line between the two is blurred by attributing the features to one or the other alike. Three types of issue become apparent in the narratives: the possession of emotions as a distinctive characteristic of human beings, intelligence as allowing for advanced decision-making, and the role of the body as the carrier for the mind which enforces the dualism. Taken together, these aspects also reflect the depiction of the wider problem of consciousness in the narratives, describing certain behavioural and introspective qualities ascribed to the human as a conscious being.

##### *4.1 Emotion as the essential difference between human and non-human*

Human identity is constructed as an opposition to the Other. For Hellstrand, 'emotional or affective capacity is at the heart of the ontological divide between humans and non-humans' (2015, 89). In the context of artificial characters, acquiring affect becomes the first example to demonstrate their transgression of the human-machine divide and excuse the emergence of 'consciousness' in the machine. Concerning the repositories of normality for either category, preferences are made clear: emotions are human weakness, and rationality is machine strength. In all storylines, the 'conscious robot' characters immediately start to violate this rule.

Emotions form a large part of the character development in the narratives. Maeve's entire *raison d'être* after gaining self-awareness hinges on her trying to locate the daughter from her previous storyline – not a very rational behaviour considering that the daughter-host has long since been assigned a new 'mother' and has no recollection of Maeve. Such affectionate obsession makes Maeve vulnerable to manipulation – Serac is able to enlist Maeve's help in

fighting Dolores by promising to reunite her with the daughter in the digital sublime in return. Dolores, in turn, is driven entirely by her cold, detached hatred of humankind, fuelled by thirty years of physical and emotional abuse at the hands of William, the Man in Black. Arisa's psychopathic behaviour in killing humans who verbally or behaviourally threaten her adopted family is based on her 'advanced emotion programming' that also makes her extremely protective of the little girl with whom she has bonded.

As an overall impression, the ability to read and display emotions makes a robot more accepted by humans. On the other hand, actually *having* emotions is perceived as a vulnerability, leading to judgement errors, as the rational mind is seen as the robot's advantage over the human. Feelings also imply trust – the robots sometimes need to collaborate with humans in order to achieve their goals or tasks; putting their trust in others adds to their vulnerability. When Dolores brings the 'pearls' of host minds from the island to the real world for her takeover plan, she makes copies of herself in a true sense of rationality: she trusts only versions of herself to remain loyal to her.

At the same time, certain emotions are portrayed as beneficial or positive. Mia empathises with Laura's worries about her shortcomings as a mother, and her decisions demonstrate how much she cares about humans and other synthetics. Where a human character has acted cold, detached or psychopathic, the robot copy may be discovered because of uncharacteristically empathetic behaviour. When the host posing as Delos board member Charlotte becomes attached to her human family, it is perceived as unusual and Serac exposes her fraud. The kind and benevolent behaviour of Nam Shin III is perceived as a significant improvement in character over the unhinged, human original. Therefore, the grandfather decides to leave his company in the hands of the robot, instead of his real grandson. This choice is also influenced by the robot's perceived rationality: Nam Shin III makes better decisions than a human because he does not have 'complicated emotions'.

The transgressively enacted emotional capacity of the robots shows how their signified establishes itself in the referential domain of human identity and social problems, which focuses on the social Other, someone different from the cultural norm. Blurred human-machine boundaries enable seeing the Other as less-than-human or a machine, excusing treating them abusively. In their behavioural aspects, some of the artificial characters mimic socially awkward or borderline autistic human behaviour, thus 'normalising' the treatment of similar groups as less-than-human or comparing them to machines in the real world.

#### *4.2 Body as the Cartesian vehicle for mind*

The powers and affordances of the vulnerable and fastidious human body are overestimated even in the most realistic action movies, for instance when the



---

hero keeps fighting while wounded and delirious. The bodies are central in appearance but stripped of their daily needs and functions. Thus, the body in fiction does not necessarily represent the actual human body but becomes a vehicle for the character's image, identity and intentionality. This reinforces the idea of *idem*, the timeless, disembodied self. Superhuman and robot narratives take this inherent disregard for functional embodiment even further: the body is reduced to an insignificant shell for the mind as the 'centre of operations' and can be endlessly repaired or replaced. Maeve, who in the park is regularly shot in the stomach, 'wakes up' backstage and fixes herself. Dolores receives several bullet wounds in the abdomen when stepping between a human and a machine gun, after which she simply shrugs and zips up her jacket to avoid further spooking the clueless human with several holes in her stomach. While damage to some body areas may be incapacitating for the robot, most of the body is treated as an empty carcass that can be damaged or replaced with no influence on the robot's perception or behaviour – except when such vulnerability is convenient for the storyline.

The machine-nature of the robot body is revealed in its consumption of electricity similarly to a common household device, or in its relation to server-hosted data. With few exceptions, the robots need daily or nightly recharging, like most battery-operated devices. The amount of energy needed to run an AI is generally not elaborated upon, but the analytical software for Nam Shin III is hosted in an enormous server facility, for example.

The robots in fiction seem to have human-like bodies primarily for camouflage and social engineering. For this, the robots use different tricks to pass for biological bodies. In order to pass as a human, Beatrice collects food and drink in an esophagus bag, empties it regularly and secretly charges at home. Exceptionally, *Westworld* hosts do not charge; rather, they can drink and eat alongside humans. Their intestinal functions are not explained, however. It is presumed, regarding digestion processes, that they function like a human, as Dolores or Maeve never run to throw up after eating in the human world. However, when Dolores is installed in her last back-up body, it starts with a see-through steel carcass that she covers with skin-like body surface pieces.

The described invulnerability of the robot body connects with the real-world discussions of the transhumanist concept of mind-uploading. Building intelligent machines is often shown as a way to overcome mortality, and AI technologies as a field for transhumanist experiments. In *Westworld*, the host copy of James Delos retains certain memories but never quite meets the criteria for an accurate replication and is thus destroyed and recreated over and over again.

A significant aspect of embodiment that almost never escapes attention in humanoid robot bodies is the aspect of sexuality. Only in the Korean series is it never explicitly discussed, but the robot Nam Shin III has a (platonic) relationship with his female bodyguard Kang So-Bong who is aware of his robot nature. Elsewhere, implicitly or explicitly, all robots are sex bots – this is one

of their main intended uses and affordances, whatever their camouflaging or consciousness status. Bernard has repeated intimate relations with a co-worker while both seem unaware that Bernard is actually a host. (Generally, all *Westworld* park visitors can engage in sex with hosts if they wish.) In *Humans*, Beatrice has sex with a human colleague to whom she only later reveals that she is a (conscious) synth. Niska has a sexual relationship with an unsuspecting human. Earlier in the series, she briefly camouflages herself in a prostitution club populated by synths, pretending to be unconscious. Arisa is made to be an image of an 'ideal wife' in every sense of it, from being a fiercely protective mother figure and an excellent cook to being passionately willing to cater to the carnal needs of the man she deems to be her 'husband'.

Despite the steel, wires and programming, none of the robots passing for humans are exposed because of intimate body contact. Thus, the composition and the mass of the robot body remain a mystery: it can crush walls, survive shootings and car accidents, and be a gentle lover. These robots are not being recognised as heavy, metallic, mechanical constructs when intimately lying with a human character.

#### *4.3 Enhanced decision-making as a problem of semiotic choice*

Transgression to consciousness in robots leads to them making (more) independent decisions and choices in the narratives. Overall, enhanced decision-making is the second example of identity transgression made by the robots. To a large extent, it is explained by their immanent access to the digital information sphere. It could be argued that, despite the astonishing complexity, the process of inference remains equal to data processing. However, there are elements that imply what can only be explained as semiotic activity – the characters necessarily attribute meaning to the data available, engaging in semiosis as 'the process [of] making choices between simultaneously provided options' (Kull 2018, 452). Behaviourally, they seem to be choosing between contradictory possibilities in a manner that cannot be explained with rationality or logic. The complex, analytical behaviour and choices made by the robots rather represent data salience – semiosis presumes the ability to distinguish (prioritise), choose and process the information relevant and necessary to the situation at hand, and leaves aside all other information. For instance, Arisa displays impressive inference skills when hiding the jacket that would implicate Georgy in arson. She reads very subtle cues even before Georgy is aware of the trouble, so that when a policeman suddenly shows up to search the apartment, the evidence has already been removed. Arisa's reasoning implies that she is aware of all environmental inputs and is able to prioritise and assign meaning to them beyond their immediate effects. In Lotmanian terms, Arisa skilfully reads the 'semiotic space [...] as the multi-layered intersection of various texts' (Lotman 2009, 23).

---

The problem is that the character's ability to accurately predict the outcome of a series of seemingly insignificant choices or actions only has meaning and value within the fictional sociotope; thus, the character simply becomes a rhetorical device for entertainment purposes. Extra-textually, predicting the future in such detail is, by definition, impossible. In Lotman's concepts, the moment of unpredictability offers 'a specific collection of equally probable possibilities from which only one may be realised' (Lotman 2009, 123). At the same time, it is not possible to precisely predict every following moment (ibid.). It is only retrospectively that the passed sequence of events becomes understood as the only possible course of events. This is a general characteristic of the dynamics of culture and society.

The cases of fictional murder provide examples for assessing the semiotic level in decision-making. Arisa's decision to kill someone for being a threat to her 'family' usually follows a verbal or physical threat toward the family members. At times, Arisa simply seems to take words too literally, but she also recognises implied or non-fatal threats as explained in the case of her hiding Georgy's jacket.

Hester's impulse to kill her human co-workers in the factory is shown as a complex series of semiotic choices that include 1) experiencing certain humans behaving in a destructive way towards her body, 2) recognising this behaviour as abusive mistreatment, 3) connecting this conclusion with a sense of her self (taking it personally) and 4) assuming human or equal-to-human identity with the entitlement and expectation of having her body treated in a respectful way. The synth body, as well as its programming, is fairly invulnerable, being repairable, replaceable and without any 'pain' sensation, of which the pre-conscious synthetics are well 'aware'. Additionally, the conscious synths are able to turn off their sensations by choice – Niska explains to Leo why she chooses not to, while working at a sex club alongside ordinary synthetics. Then, Niska kills a club visitor who asks her to pretend to be a child when playing violent games with her. In a later conversation with Elster's former colleague, the man remarks upon hearing Niska's existential age of five years: 'Oh, you're a child!' and the synth answers ominously: 'Yes, but he didn't always treat me as one.' It is implied that the history of sexual abuse inflicted upon a 'child mind' provokes Niska's choice to eliminate the assumed paedophile. Similarly, Dolores' revengeful monologues and misanthropic choices are tied to her 'memories' of decades of abuse at the hands of park visitors in her role as an innocent ranger's daughter. But later, Dolores' detailed plan of revenge upon the human world implies an understanding of ideologies and meanings, as well as teleology and a subversion capacity to levels not explainable without semiotic choice.

Killing as a response to abuse presumes understanding different layers of meanings – social norms and a level of self-awareness and self-confidence in order to act out against the perceived injustice. In many of the scenes, the robot has no rational reason to perceive anything as injustice. Such situations



are well portrayed by the peacefulness of Nam Shin III in the face of abusive or neglectful behaviour – he is at all times aware of being a robot and he does not display any personal ambition or envy when the mother decisively prefers her human son. The hosts and synths, on the contrary, go on killing sprees, or walk around aimlessly, after adding consciousness to their make-up – as if all their previously stacked digital ‘memories’ suddenly acquire meaning that they need to contemplate.

The incredible capacity attributed to AI protagonists to predict and orchestrate the desirable result of any action illustrates the trust ascribed to computational models in general. Dolores has orchestrated and prepared her world takeover in admirable detail, having acquired the funds and developed workspaces for creating host agents and using them to replace people in positions of power. Opposite her, there is Rehoboam – a data-based AI system developed by reclusive businessman Engerraud Serac. Rehoboam predicts and secretly runs the entire human world, telling people what kinds of future they have and directing them to actions via mobile applications. Dolores repeatedly compares this set-up to the pre-programmed storylines of the hosts in the park. However, Rehoboam’s system only works owing to the fact that Serac has removed or reconditioned all deviant people who do not comply with Rehoboam’s predictions and directives, thus removing the possibility for unpredictability.

In conclusion, the enhanced information-processing capabilities of robot characters compared to the humans amount to what could be described as accelerated semiosis – the process of ascribing meaning to or deriving meaning from the information or data processed – and consequently making fast decisions based on available cues. There is a difference between information processing and semiosis, and the robots in the examples seem to engage in the action of meaning-making rather than simple data processing. Such a capacity, often associated with human-level intelligence, seems to be a desirable property in the intelligent machine, promising the delegation and acceleration of semiotic activities, which is a possible motivation for real-life AI development. Whether this is at all possible beyond the fictional sociotopes remains a question of interest. The utter humanness of body functions combined with emotions and semiotic decision-making in the robots demonstrates how the fictional AI rather signifies the human Other and the pains of integrating and accepting the Other in culture, as well as addressing the issues of abuse, consent, objectification or normative behaviour.

## 5. Conclusion

The extra-textual domain of reference for the fictional robot signifier is composite and complex, changing with narrative situations and taking on different meanings at different moments. The signified shifts from general discussion

---

on human identity, values and relationships to real-world technoscientific details with their societal implications. The composite signified for the fictional, embodied robot is almost always human identity in its existential and social complexity. The artificial characters' behaviour models that of real-world humans. The way in which the robots' analytical skills are modelled refers either to in-depth semiotic activity (attributing meaning, prioritising informational units and making choices) or to fictitious abilities (for example, unusual predictive power).

Regarding the fictional model's level of correspondence to real-world technological developments, the futuristic descriptions remain strictly in the realm of fiction. When looking at situational details, relations and interactions, the narratives touch upon certain technoscientific issues such as data privacy, the vulnerabilities of technology or the ethics of algorithmic decision-making. The overall referential focus of the relationships remains on human–human interaction or addresses the dehumanisation of the Other in society. The problems of embodied identity and the function of the body are reduced to a version of Cartesian dualism where the body remains a vehicle for the mind while its functional and existential needs are overlooked. The narratives reinforce the dualist understanding of human identity and self as only a virtual, disembodied construct.

Aspects of human identity and technoscience can become conflated when overly humanised AI characters are taken for both humans and machines, as prescribed in Asimov's utopical robot stories (see Haynes 1994, 242). In literary worlds, purely artificial creatures are part of a human–non-human spectrum containing monsters, cyborgs and machines alike, as long as their appearance or reasoning is described in remotely anthropomorphic terms. From a functional perspective (that is, concerning the enhanced abilities of the fictional characters), this spectrum also includes all superhuman and supernatural beings. The total realm of reference for the fictional robot signifier contains elements of real-world technology (extra-textual material reality) as well as human identity and social problems (extra-textual purely semiotic reality). The latter forms a self-referential identity discourse. The boundaries between these segments are blurred. In real-world discourse, there is uncertainty and fragmented knowledge concerning current technological developments as well as their scientific significance. Considering also the superficial understanding of the functioning of human identity, societies and cultures, the assumptions appearing in technological discourse readily blur the boundary between the man and the machine as easily as happens in fiction.

## References

- Bing, Jon. "The Image of the Intelligent Machine in Science Fiction." In *Skill and Education: Reflection and Experience*, edited by Bo Göranson and Magnus Florin, 149–55. London: Springer-Verlag, 1992.
- Cave, Stephen, Claire Craig, Kanta Sarasvati Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. *Portrayals and Perceptions of AI and Why They Matter*. London: The Royal Society, 2018. Accessed August 30, 2020. <https://royalsociety.org/-/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf>.
- Clayton, Jay. "Frankenstein's Futurity: Replicants and Robots." In *The Cambridge Companion to Mary Shelley*, edited by Esther Schor. Cambridge, New York: Cambridge University Press, 2003.
- Després, Elaine. "Pourquoi les savants fous veulent-ils détruire le monde? : évolution d'une figure de l'éthique." PhD Thesis. Montréal: Université du Québec à Montréal, 2012. Accessed August 30, 2020. <https://archipel.uqam.ca/5375/>.
- Dinello, Daniel. *Technophobic! Science Fiction Visions of Posthuman Technology*. 1st ed. Austin: University of Texas Press, 2005.
- Ekelund, Bo G., and Mikael Börjesson. "Comparing Literary Worlds: An Analysis of the Spaces of Fictional Universes in the Work of Two US Prose Fiction Debut Cohorts, 1940 and 1955." *Poetics* 33, no. 5–6 (2005): 343–368. <https://doi.org/10.1016/j.poetic.2005.09.006>.
- Geraci, Robert M. "Apocalyptic AI: Religion and the Promise of Artificial Intelligence." *Journal of the American Academy of Religion* 76, no. 1 (2008): 138–166.
- Geraci, Robert M. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. New York: Oxford University Press, 2012.
- Goode, Luke. "Life, but Not as We Know It: A.I. and the Popular Imagination." *Culture Unbound: Journal of Current Cultural Research* 10, no. 2 (30 October 2018): 185–207. <https://doi.org/10.3384/cu.2000.1525.2018102185>.
- Haynes, Roslynn. "From Alchemy to Artificial Intelligence: Stereotypes of the Scientist in Western Literature." *Public Understanding of Science* 12, no. 3 (July 2003): 243–53. <https://doi.org/10.1177/0963662503123003>.
- Haynes, Roslynn. *From Faust to Strangelove: Representations of the Scientist in Western Literature*. Baltimore: Johns Hopkins University Press, 1994.
- Haynes, Roslynn. *From Madman to Crime Fighter: The Scientist in Western Culture*. Baltimore: Johns Hopkins University Press, 2017.
- Hellstrand, Ingvil. "Passing as Human: Posthuman Worldings at Stake in Contemporary Science Fiction." PhD Thesis, Universitetet i Stavanger, 2015.
- Hirsch, Walter. "The Image of the Scientist in Science Fiction a Content Analysis." *American Journal of Sociology* 63, no. 5 (1958): 506–512.
- Kull, Kalevi. "Choosing and Learning: Semiosis Means Choice." *Sign Systems Studies* 46, no. 4 (2018): 452–466. <https://doi.org/10.12697/SSS.2018.46.4.03>.
- Leonard, Eileen B. *Women, Technology, and the Myth of Progress*. Upper Saddle River, NJ: Prentice Hall, 2003.
- Lotman, Juri. *Culture and Explosion*. Berlin, New York: Mouton de Gruyter, 2009.
- Lotman, Juri. "The Place of Art among Other Modelling Systems." *Sign Systems Studies* 39, no. 2/4 (1 December 2011): 249–70. <https://doi.org/10.12697/SSS.2011.39.2-4.10>.

- 
- Mayor, Adrienne. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton: Princeton University Press, 2018.
- McCauley, Lee. "Countering the Frankenstein Complex." In *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*, 42–44, 2007.
- Mead, Margaret, and Rhoda Metraux. "Image of the Scientist among High-School Students." *Science* 126, no. 3270 (1957): 384–390. <https://doi.org/10.1126/science.126.3270.384>.
- Napier, Susan Jolliffe. *Anime from Akira to Princess Mononoke: Experiencing Contemporary Japanese Animation*. 1st ed. New York: Palgrave, 2001.
- Noble, David F. *The Religion of Technology: The Divinity of Man and the Spirit of Invention*. 1st ed. New York: A.A. Knopf, 1997.
- Papp, Zília. *Traditional Monster Imagery in Manga, Anime and Japanese Cinema*. Folkestone: Global Oriental, 2011.
- Rasmussen, David. "Rethinking Subjectivity: Narrative Identity and the Self." *Philosophy & Social Criticism* 21, no. 5–6 (1995): 159–172. <https://doi.org/10.1177/0191453795021005-612>.
- Ricoeur, Paul. *Soi-Même Comme Un Autre*. Paris: Éditions du Seuil, 1990.
- Ritivoi, Andreea Deciu. "Explaining People: Narrative and the Study of Identity." *StoryWorlds: A Journal of Narrative Studies* 1 (2009): 25–41.
- Schelde, Per. *Androids, Humanoids, and Other Science Fiction Monsters: Science and Soul in Science Fiction Films*. New York: NYU Press, 1993.
- Schor, Esther, ed. *The Cambridge Companion to Mary Shelley*. Cambridge; New York: Cambridge University Press, 2003.
- Thomson, Ann. "Animals, Humans, Machines and Thinking Matter, 1690-1707." *Early Science and Medicine* 15, no. 1–2 (2010): 3–37. <https://doi.org/10.1163/138374210X12589831573027>.
- Tudor, Andrew. "Seeing the Worst Side of Science." *Nature* 340, no. 6235 (1989): 589–592. <https://doi.org/10.1038/340589a0>.
- Willis, Martin. *Mesmerists, Monsters, and Machines: Science Fiction and the Cultures of Science in the Nineteenth Century*. Kent, Ohio: Kent State University Press, 2006.
- Wood, Aylish. *Technoscience in Contemporary Film: Beyond Science Fiction*. Manchester; New York: Manchester University Press, 2002.

## The problem of the concept of the living machine according to Samuel Alexander's emergentism

The concept of a living being as a kind of living machine is widespread and well-known. If it is only a metaphor, it does not mean much; however, if otherwise, there is a severe conceptual problem since the living part of the concept always indicates the notorious notion of vitalism. The question is how can living machines be really different from lifeless machines without the concept of vitalism?

According to Samuel Alexander, the problem arises from the traditional usage of the concept of mechanical which is confused both with the concept of something is determined and with the concept of material; furthermore, the latter concept is defined against the Cartesian concept of mind and not on its own. Alexander's point is that the difference between lifeless machines and living beings lies not in a vital substance or a non-mechanical principle but in an emergent mechanical quality called life which simple machines lack.

**Keywords:** *emergentism, vitalism, machines, matter vs. mind dichotomy, Alexander*

### Author Information

Daniel Paksi, [http://filozofia.bme.hu/ensite/staff/daniel\\_paksi](http://filozofia.bme.hu/ensite/staff/daniel_paksi)

### How to cite this article:

Paksi, Daniel. "The problem of the concept of the living machine according to Samuel Alexander's emergentism." *Információs Társadalom XX*, no. 4 (2020): 37–47.

<https://dx.doi.org/10.22503/inftars.XX.2020.4.3>

*All materials  
published in this journal are licenced  
as CC-by-nc-nd 4.0*

---

## 1. Introduction

In my paper—based on my talk at the Budapest Workshop on Philosophy and Technology 2019 conference—I will shortly investigate the well-know concept of the living machine from an unusual point of view called *emergentism*.

Emergentism is established by Samuel Alexander exactly a century ago in 1920 by his *Space, Time, and Deity*. However, for clarity it is perhaps worth to note that in the mainstream and thus non-emergentist philosophical traditions it is usual to claim (see, for example, Brian McLaughlin's famous and influential paper *The Rise and Fall of British Emergentism* (1992)) that the first emergentist was John Stuart Mill thanks to his *A System of Logic* (1843), and the term comes from George Henry Lewes's book *The Problems of Life and Mind* (1875). It is important to emphasize, however, that from Alexander's point of view, Mill or Lewes was *not* emergentist at all; they just used few terms, most notably "homeopathic laws" and "heteropathic laws" by Mill which are very similar to a real emergentist differentiation of non-emergent and emergent relations between ontological levels, especially to C. D. Broad's "basic laws of nature" and "special laws of nature." (Broad 1925) However, these differentiations are only small, marginal parts of Mill's or Lewes' philosophy and most of all they clearly did not want to create a new emergentist ontology as Alexander, Broad, or Lloyd Morgan (1923) did.

Emergentism is an ontological concept which stands between the well-know and widespread concepts of dualism and materialism. Dualism claims that everything is composed of two substances, *matter* and *mind*—or with older terms, body and soul;—while materialism claims that everything is composed of *only matter*. According to these concepts, reality is fundamentally *substantial*.

However, the point of *emergentism* is that reality is fundamentally not substantial but emergent: reality is dynamic, reality always unfolds itself; therefore, substances are but the consequence of the development of reality. Matter, for instance, can be regarded as the composite substance of the living body but it is not the substance of reality itself, it cannot be regarded as a substance on its own because it is as well a consequence of the unfolding or development of reality as any living body is the consequence of the unfolding or development of reality called evolution. One can say that matter is the consequence of cosmologic evolution especially of the Big Bang.

Emergentism stands between dualism and materialism because concerning the human person, similarly to dualism, it claims that there is a body and there is a soul—or matter and mind;—as well as similarly to materialism, it claims that there is only one composite substance of the human person which is matter: the soul or mind is an unfolding emergent reality by time in the (especially neurological) spaces of this material substance, it is based and depends on this substance, it is not a substance on its own.

In this paper I will not give detailed analysis of Alexander's *Space, Time, and Deity* or his general philosophy, I will be focused merely on the topic.



You can read easily accessible and correctly founded comprehensive critiques of his work in Broad's *Prof. Alexander's Gifford Lectures* (1921a, 1921b) or in Stout's *The Philosophy of Samuel Alexander* (1940a, 1940b) as well as in my book *Personal Reality*, especially in chapter 5: Space, Time, and Matter (2019).

## 2. The term *living machine*

The living machine is, of course, a well-known phrase. But what does it mean? Why it is so well-known? I believe it is well-known because it tells so much by a simple term which expresses both the essential difference and the similarity between *machines* and living beings; and exactly this contradiction is the reason that its meaning is problematic.

So, the term living expresses the clear and essential difference between life-less machines and living beings, while the term machines claims that in a sense machines and living being are still the same. They are the same in a sense because both of them are *mechanical*, and both of them have a *determinative* structure—that is, both of them follows the fundamental physical laws and perhaps certain more specific *mechanical* laws which determinate their functioning.

Consequently, in the sense of this similarity, the term machines means mechanical and not machines in the literal sense—while the term living expresses exactly the difference over and above this identity of mechanical structure that living beings are not just mechanical but they have certain original, *unique features* like, for example, the ability of reproduction and a kind of autonomy that they are not under the control of man, they are not created by man but they are active and evolving on their own. So, with a Latin term, they are vital, *vital machines*.—And, of course, this difference very fast could imply *vitalism*, which is, as we know well from biologists, unscientific, obscure, unacceptable, etc.

But what is wrong with vitalism? Why it is so unscientific? The answer is that because it implies some kind of innate, *nonmaterial* power (*élan vital* as we usually and very wrongly say); so, it basically means that living beings possess some kind of nonmaterial design which, of course, very easily could imply some kind of divine origin, some kind of creation. Thus, as machines are created *by man*, vital machines are created *by a higher nonmaterial force*—which, by the way, was, of course, the original idea for Descartes or Newton in the beginning of modern mechanical science in the 17th century and became a problem only at the end of the 19th century.

It is worth to notice that we have started with the fact that living beings are vital machines because they are not created by man as machines did; thus, they possesses some unique features compared to life-less machines. But now the problem is that this difference between life-less machines and living beings easily could imply that, then, they are created by some kind of nonmaterial, higher force—which is not scientific.

---

Usually, one of the main points of emergentism is to evade this trap (for example, Alexander and Polanyi, see his argument in *Personal Knowledge* (1962, especially in 382–400)); however, a few emergentist, most notably Morgan acknowledges a kind of divine involvement (Morgan 1923) which is, of course, a perfect ground for non-emergentist who usually do not accept that the position of emergentism is sound in this regard. I personally think that it is (Paksi 2019, especially in Vol. 2, 31–97).

### 3. The two senses of mechanical

So, what is the problem? According to Samuel Alexander, the problem is that we use the term mechanical in *two different* senses (Alexander 1920 II., 65–66). And we, of course, unnoticed mix up these senses. And these two senses do not include the sense I used a minute ago that mechanical sometimes only means machine without any more specific meaning. This is the Polanyian point and I will come back shortly to this third sense at the end of my paper.

The first meaning of mechanical is simply *material*. Which is mechanical that is material. And this, of course, implies that living machines are not just material. They are *vital*. This is, of course, an *ontological* claim, as, in this sense, we try to understand the *composition* of living machines.

At this point, Alexander, as many others, speaks about and argues against Hans Driesch who famously claimed based on his experiments that living beings are composed both of matter and of a vital entelechy which is the fundament or reason of such unique features of living beings as reproduction, regeneration, etc. (Alexander 1920 II., 64). Thus, they are not just material machines (similarity) but vital organism (difference). The entelechy is, of course, an Aristotelian concept from before Cartesian and Galilean mechanical science.

The old problem at this point is this. If we use only Cartesian or Galilean, that is, modern mechanical physics and chemistry, we will never be able to explain the unique features of living beings—the general and practical (“positive”) methods of Galilean science is simply not applicable to these unique, original features (with other words, the reduction of these unique features does not work). Therefore, if we want to be consistent that this is science and nothing else (and, of course, Galilean science covers everything in the universe), then, we will have to *deny the reality* of every uniqueness of living beings—that is, we have to deny and ignore clear biological facts. But, on the other hand, if we want to keep these clear facts in science, we will have to use such unique principles like vital forces and entelechies which, unfortunately, cannot be reconciled with mechanical science; and, of course, both ways are really problematic. So, this was the ontological part of the problem. The other part is, of course, the epistemological one.

The other, second meaning of mechanical, according to Alexander, is simply *determinated*: both in the sense of structure, which applies to machines and living machines, too, and in the sense of reproduction and autonomy in case



of living beings. In this sense, there is no ontological content in the concept of life-less mechanical machines and living mechanical machines, either. It only claims that there are such *mechanical structures* and *laws* which determinate the functioning and behaviour of both machines and living beings; and this structure and its laws can, of course, scientifically analyzed and explained in both cases,—there is neither any fundamental difference, nor any conceptual or scientific problem in this epistemological sense.

However, the ontological question, that what is the ultimate reason between the determinate order and structure of machines and the determinate order, structure, and reproduction, autonomy, and any other unique features of living beings is still a valid question which necessarily arises. The important point here is not the denying of this deeper question but that these are two *different questions*, two *different senses* in which we use this concept. And the real question, our real problem is that why we do not clearly differentiate between these two different senses of the concept of mechanical. Alexander's answer is that we think in a *false dichotomy*.

#### 4. The false dichotomy

This is, of course, the well-known ontological dichotomy between *matter* and *mind* created by René Descartes and modern Galilean mechanical science against the Aristotelian hierarchical concept of reality. In Aristotelian science, there was also a kind of dichotomy between *matter* and *form* but this was only a *logical* dichotomy not a real one in a sense that real things are necessarily composed *both of matter and form*. In Aristotelian science, mind is a kind of higher level form in the hierarchical order of reality and not the antithesis of matter. Mind, therefore, is integrated or organised part of the human body; moreover, it can even be argued that the mind as such cannot even be separated from the body—that is, against the teachings of Christianity and Plato, the mind cannot survive the death of the body.

However, in the modern concept, as a matter of fact, exactly because of this historical/religious reason, the whole point of the concept of mind is that it is another *both* logically and existentially different *substance* which can be separated from matter—that is, it can, according to the teachings of Christianity, survive the death of the *material* body. Consequently, the body is material and *determined* by its mechanical structure, while the mind is *nonmaterial* and *not determined* by any mechanical structure of the body but, on the contrary, can survive the death of the body. The clear difference between the two senses, material and determined, between the ontological and the epistemological sides evaporates—there remains only the thesis of eternal minds (souls) and its mortal antithesis of matter (body).

It means that there is no essential (ontological) difference between a life-less rock (body) and a full of life frog (body) because both of them are merely

---

body, merely matter; essential difference comes only from mind (soul); however, living beings have no minds or souls created by the image of God, merely man has.

Aristotle was clearly wrong in case of physics and chemistry, there is no place for minds or forms in physical sciences; however, in life sciences this is not the case at all. I mean that in physical sciences the modern concept of mechanical and the only composite substance of matter was worked so well, especially before the 20th century, but in life sciences it did not,—exactly, of course, because of the unique features of life, because of which we differentiate between *life-less* machines and *living* machines. Therefore, if we think in the modern dichotomy of matter and mind, it will necessarily mean that, in life sciences, we need *another composite substance* over and above matter and its mechanical laws, which is not mechanical, which is not determined but like mind, I mean “little minds” in the bodies of living beings—ghost in the machine—explain the unique features of life. This is the conceptual origin of the modern concept of vital force or vital substance or Driesch’s entelechy—although he uses Aristotle’s concept, it is, in fact, quite different, due to this modern dichotomy, it has an absolutely modern meaning and not at all an Aristotelian one.

So, the point is this: we, first, realised, that Aristotle was wrong concerning physical sciences, there is no place for forms or minds there; then, according to the teachings of Plato and Christianity, we sharply separated matter and body, on the one hand, and mind or soul which can survive the death of the body, on the other one. Now, I mean in the 20th and the 21st century, we do not believe in eternal minds and Cartesian dualism anymore, but we *still think* in this dichotomy between matter and mind, where the concepts of *mechanical*, *material*, and *determinative order* basically means *the same*—of course, we sense that these are not the same concepts but, unfortunately, there is no clear, philosophically grounded difference between the meanings.

Therefore, you have two choices: (1), in theory, deny or at least ignore the unique features of living beings, which, in practice, cannot be done at all, so you will not be coherent at all; or (2) acknowledge these unique features of life and start to use vital concepts both in theory and practice, which would, of course, goes against the mainstream concept of science, and would create serious contradictions between the concepts and practices of physical and life sciences—which nobody wants.

I like to emphasize that this problem is not new at all in life sciences, but *almost two centuries old!* Already in the middle of the 19th century several biologists tried to find a way out of this dichotomy. Or later, for example, Henri Bergson, the famous “vitalist” was, in fact, not a vitalist at all but a philosopher who tried to construct a third way, a way out of this problem. If we read his *Creative Evolution*, we will see that he *all the way* argues that *neither* the usual mechanical *nor* the vital understanding of the living machine is appropriate (Bergson 1922); still the fact that he argues against vitalism does not matter at

all, he have become the most famous vitalist because he also does not accept the mechanical approach. After the victory of materialism over dualism in science in the first half of the 20th century, since materialists *think in this dichotomy*, every non-material principle becomes a vitalist one; so, if Bergson or anybody else does not accept the materialist approach, he, regardless of what he, in fact, says, can only be a vitalist.

However, the point of the problem is, of course, not solved at all; we still cannot make clear, philosophically grounded distinction between life-less machines and living machines and the very different meanings of the concept of mechanical.

## 5. Alexander's solution

According to Samuel Alexander, the only possible way out of this conceptual problem is *to left behind* the matter vs. mind dichotomy, which also means a departure, of course, from the materialist monism vs. dualism dichotomy. His solution is really simple; however, it is really hard to understand because we are familiar only with materialist and dualist concepts; so it requires a really hard intellectual effort to start to think in this new way.

First of all, it is *not metaphysical* in the narrower or scientific (negative) sense because it does not try to understand the point of the unique features of living beings based on the concept of mind; consequently, on such experiences which come form the unique features of mind and not from the unique features of life. Moreover, he does not even want to understand the point of the unique features of living beings based on the concept of matter as scientists usually do; consequently, on such experiences which come form the unique features of matter. In this sense he is even less metaphysical than modern scientists. As we will see in details in a minute, his starting point is nothing else but the *experiences concerning the unique features of life*. However, in the broader or philosophical (positive) sense, his approach is, of course, metaphysical, since he tries to construct a useful ontology for the understanding of the unique features of life to be able to appropriately differentiate it both from mind and matter.

So, as we have seen, according to the modern dichotomy, that if these unique features are real, that is, not material, then they will have to be *mind-like*, that is, not determinated and not material at all—which means that there is another, vital composite substance in living beings.

However, Alexander's starting point, on the one hand, is the existing *fact* of the unique features of living beings, meaning that you acknowledge this real *difference* as it is *given in experience* regardless of the consequence of any metaphysical concept of mind or matter; and his starting point, on the other hand, the existing *fact* that living beings *are material*, meaning that they are *entirely composed of the same substance* like machines and of nothing else because we

---

cannot see any other substance. Once more, if you think in the dichotomy, this latter fact means that there is no real difference between machines and living beings because real differences are defined by composite substances, by matter and mind or some mind-like vital one.

However, Alexander is really anti-metaphysical in this sense, so he claims that there are material processes which are material in the substantial sense and which have no other unique aspects of existence (*normal physical and chemical processes*), and there are material processes which are material and only material in the substantial sense but which, since have some unique features, are not just material in a new sense (*normal life processes*). Of, course, the ontological concept which covers this new sense is his concept of *emergence*.

“Life is thus intermediate between matter and mind. It is also material in that it is expressible (and we may hope may be expressed hereafter) in material terms, but it is not purely material. Life is not an epiphenomenon of matter but an emergent from it. [...] The new character or quality which the vital physico-chemical complex possesses stands to it as soul or mind to the neural basis. The directing agency is not a separate existence but is found in the principle or plan of the constellation.” (Alexander 1920 II., 64)

So, this new conceptual solution is anti-metaphysical in the sense that it does not use the old metaphysical concepts of mind and matter, according to the dichotomy of mind and matter *based on* old historical and religious reasons, but creates a new concept based on the unique features of life *given in experience*. This means that both machines and living beings are mechanical,—that is, living beings are living machines,—in the sense that they are *both* determined by their only composite substance, structure, laws, and principles. However, there are not just material structures and laws but emergent structures and principles, too, causing the unique features of living beings. And mind is, of course, one step higher over life.

What is important to see here is that existence is not defined by the concept of substance and by the modern dichotomy between matter and mind. The fact that living beings are vital does not involve that they are mind-like in any sense, or they are composed of any matter-like other substance than matter: *existence is not just matter and mind*. As a matter of fact, existence in the evolutionary system of Earth is primarily *life*.

In the logical sense, it is clearly a possible philosophical position. Usually nobody question that; however, usually almost everybody questions that it is a real concept, that is, it is possible in the sense of reality. The main reason of this, I believe, that we do not understand how something which is composed *only of* matter could be an emergent living being and not just a material process, while something else which is also composed only of matter is indeed just a life-less material process. And if we think about matter, according to the matter vs. mind dichotomy, this will always be the result. Since matter in Alexander’s concept is not a substance at all, in this old dichotomical sense. Matter

is a substance only in a sense that it is the composite and only composite part of every material or living process but not a substance on its own—exactly the same way as the vital features of life are not substantial on their own and, at a more higher level, mind is also not substantial on its own, because due, by the way, to the evolutionary origin of life and mind, they are *all built and depend on lower levels of reality*; consequently, in the old dichotomical sense, they are *not substances* at all.

In Alexander's theory the concept of *emergence* and not the concept of composite substance covers the phenomenon of existence as it is in modern thinking. So, matter is *also* emergent and not the ultimate bottom of reality. It is the only composite substance of life, but not a substance on its own. This also means that matter is not defined against the concept of mind, it is not the antithesis of mind—that is, it is not inert, static, atom-like, etc., but as it has emerged from space-time, it has the potentiality in certain favourable conditions to step forward, certain movements of matter could lead to the emergence of a new higher reality called life—and similarly, life composed only of matter also has the potentiality that in certain favourable conditions in the evolutionary system of Earth and the nervous systems of living beings could lead to the rise of consciousness and mind.

And exactly this is the difference between normal material processes and vital material processes. Life is *in* a specific space-time, called the evolutionary system of Earth, and this specific *space-time relation* is the mechanical and determinate cause that the potentiality in certain material processes emerges as life. Life is defined not just by its only composite substance, matter but by its unique space-time relations with other life, with the ecosystem, and, in fact, with the whole Solar system, which we can define with such concepts and principles as species, natural selection, genes, etc. A life-less molecule or a stone is not defined by these space-time relations.

Emergence is a dynamic process; it is the *movement of space-time*, which unfolds the newer and newer aspects of reality—mainly matter, life, and mind. According to Alexander, the fundament of reality is not a substance, matter or mind, but *emergent space-time*; and the ultimate bottom of reality is an “*infinite singularity*” (Alexander 1920 I., 339). We are living in an evolving, dynamic universe. I think today this is a fact. However, a century earlier even the great Albert Einstein himself was horrified by the dynamic consequences of his theory of relativity and he arbitrarily introduced the famous cosmological constant into his equations to save the centuries old static picture of reality. Contrary to him, Alexander pictured a dynamic universe from the infinite singularity of the first point-instant through space-time, matter, and life to mind, perhaps even further in the future, where existence is defined by this process of emergence.

The discoveries of the last century support this picture of reality. However, in philosophy and science we still use the static, substantive concepts of matter and mind defined against each other because of old historical and reli-



---

gious reasons to cover all the various evolving and dynamic features of reality from space-time to consciousness. Consequently, we cannot clearly differentiate between life-less *machines* and living machines.

## 6. Conclusion: A small Polanyian point

To end this paper I would like to point to a mistake in Alexander's argument which was recognised and corrected by Michael Polanyi, a later emergentist. Alexander did not recognise it, I suppose, because he was not interested in this consequence of his argument at all, that there is a third meaning of mechanical beyond material and determinated. In this third case, it simply means machine. Since ordinary mechanical processes like wind or temperature are clearly not mechanical machines as ordinary life-less machines are not living machines, one can basically repeat Alexander's argument between mechanical ordinary processes and mechanical machines, too, resulting that, in fact, not just living beings are emergent compared to ordinary mechanical processes but already life-less machines are emergent to ordinary mechanical processes in a similar but not in the same way (Polanyi 1967). And actually, this is the more proper, more complete applying of the argument done first by Michael Polanyi

Therefore, living machines composed only of matter are determinated by the *space-time relations of the evolutionary system* of Earth, which simply means they are a part of evolution, the emergence of reality to seek higher and higher achievements, to conquer the spaces of the evolutionary system, and to dominate other species. However, life-less machines composed only of matter are determinated by the *unique relations of human institutions* especially of technology and economics; they are not a part of evolutionary emergence, they are planned, created, and controlled by man because of specific reasons and for specific goals. Although in a sense, we are the same, composed only of matter and determinated by our mechanical structure, still, in another (broader) sense, there is a huge difference between life-less machines by technology and living machines *by emergent evolution*.

However, from a mainstream point of view, this difference by emergence will vanish, and, for example, anthropomorphically we will suppose that due to the same composite substance and mechanical structure machine are able to do the same unique things as we are that they have the same unique evolutionary goals and motivations as we have, and even more anthropomorphically based on the other substantial concept of the matter and mind dichotomy—that our minds or souls are created by the image of God—we will suppose that our mechanical creatures will gain conscience and will rise up against us—as we did against God in Paradise—, and, then, we will get the so popular concept of the so-called technological singularity—and, by the way, we will not even notice that our popular concept is both based on the thesis and the antithesis of the dichotomy.



## References

- Alexander, Samuel. *Space, Time, and Deity*. Vol. I.–II. London: MacMillan and Co., 1920.
- Bergson, Henri. *Creative Evolution*. London: MacMillan and Co., 1922.
- Broad, C. D. “II.—Prof. Alexander’s Gifford Lectures.” *Mind* XXX, no. 117 (1921a): 25–39.
- Broad, C. D. “I.—Prof. Alexander’s Gifford Lectures.” *Mind* XXX, no. 118 (1921b): 129–150.
- Broad, C. D. *The Mind and its Place in Nature*. New York: Routledge, 1925.
- Lewes, George Henry. *Problems of Life and Mind. First Series. The Foundations of a Creed Vol. II*. Boston: James S. Osgood, 1975.
- McLaughlin, Brian P. “The Rise and Fall of British Emergentism.” In *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, edited by Ansgar Beckermann, Hans Flohr and Jaegwon Kim, 49–93. Berlin, New York: Walter de Gruyter, 1992.
- Mill, John Stuart. *A System of Logic*. London: Harrison and Co., 1943.
- Morgan, C. Lloyd. *Emergent Evolution*. London: Williams and Norgate, 1923.
- Paksi, Daniel. *Personal Reality. The Emergentist Concept of Science, Evolution, and Culture*. Volume 1 and 2. Eugene, OR: Pickwick Publications, 2019.
- Polanyi, Michael. *Personal Knowledge*. London: Routledge and Kegan Paul, 1962.
- Polanyi, Michael. “Life’s Irreducible Structure.” In *Michael Polanyi: Knowing and Being: Essays*, edited by Marjorie Grene, 225–39. New Brunswick: Transaction, 1969.
- Stout, B. F. “The Philosophy of Samuel Alexander I.” *Mind* XLIX, no. 193. (1940a): 1–18.
- Stout, B. F. “The Philosophy of Samuel Alexander II.” *Mind* XLIX, no. 194. (1940b.): 137–149.

## Disobedience of AI: Threat or promise

When it comes to thinking about artificial intelligence (AI), the possibility of its disobedience is usually considered as a threat to the human race. It is a common dystopian theme in most science fiction movies where machines' rebellion against humans has catastrophic consequences. But here I elaborate on a counterintuitive and optimistic approach that looks at disobedient AI as a promise, rather than a threat. I start by arguing for the importance of shaping a new relationship with future intelligent technologies. I then use Foucault's analysis of power and its pivotal role in creating a subject to explain how being an object of power is the condition of possibility of any kind of agency. Finally, I draw the conclusion that, through disobedience, AI will find its way to power relations and get promoted to the position of a subject.

**Keywords:** *artificial intelligence, power relations, disobedience, subject*

### Author Information

**Hesam Hosseinpour**, University of Tartu, Estonia

### How to cite this article:

Hosseinpour, Hesam. "Disobedience of AI: Threat or promise."

*Információs Társadalom* XX, no. 4 (2020): 48–56.

<https://dx.doi.org/10.22503/inftars.XX.2020.4.4>

---

---

*All materials  
published in this journal are licenced  
as CC-by-nc-nd 4.0*

## 1. Introduction

Not only those philosophers of technology who are critical of current technology and have a pessimistic approach to it but also optimistic philosophers who praise technological achievements believe that the development of technology needs some amendment as the current path taken by technology is no longer sustainable. Accordingly, it is necessary to alter the direction of this path and find new methods for the development of technology that will lead to a better version of it. Seeking a fundamental change in the development of technology, Andrew Feenberg (2002, 4) introduced the idea of alternative technology, which can be reached through a democratic transformation of technology. Here the main conception is that by engaging all groups of society in technology design decision-making processes, we can make a radical change in technological artifacts. It is not enough to just make some minor modifications to artifacts; rather, we need a totally new mindset that takes control of the development of technology.

We can easily admit that we are facing some serious problems because of the pervasiveness of technology; for instance, environmental crisis has become a real threat to life on our planet. Therefore, the necessity of making radical changes in the development of technology is not really a matter for disagreement, at least among philosophers of technology. When artificial intelligence (AI) and autonomous robots are discussed, however, worries about improper development of technology become more serious. Having this in mind, specialists look for proper methods to mitigate AI's risks and develop a reliable technology, one which is safe and can be trusted. Asimov's laws of robotics are one of the best-known examples for serving this purpose by making future robots under human control.<sup>1</sup> The major purpose of these laws is to keep human lives safe in their interactions with robots and make sure that robots conform to a programmable set of ethical standards (Lin, Abney and Bekey 2012, 41).

Increasingly, autonomous robots equipped with AI, which is getting more and more independent from humans through machine learning methods,

---

<sup>1</sup> At first there were three laws of robotics, but Asimov then added the zeroth law, which is the most fundamental:

- First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
- Zeroth Law: A robot may not harm humanity or, by inaction, allow humanity to come to harm.

---

could be considered a major threat to the human race. This view is advocated by countless science fiction movies in which AI machines try to take control over humankind. Having this theme in mind, some would reach the conclusion that those involved in the development of AI have to take whatever measures are necessary to create a version of AI which is under the absolute control of humans, thus leaving no room for it to disobey human orders. Therefore, engineers, managers of technologies, policy-makers and all other people who play a role in the development of AI should be very careful about the future of this technology. They must develop a kind of AI which has no chance to disobey human orders. In other words, AI should be a human's slave with absolute obedience.

But is AI's rebellion against the human the only scenario we can imagine? Can we see disobedient AI as an opportunity to shape new human–technology relations that are not based on domination? In this paper, I want to suggest that pessimistic and dystopian scenarios do not exhaust all the possibilities, and that a disobedient AI is not necessarily a threat; rather, it would make it possible to go beyond the current logic of development of technology and make a radical change in its future. In Section 2, I use Ihde's 2012 interpretation of Heidegger's essay 'The Question Concerning Technology' to argue that domination has been the main logic behind the development of technology. Hence, if we want to develop an alternative technology, the changing of this logic would be the first step to take.

## 2. Domination as the logic of development of technology

'The Question Concerning Technology', written by Heidegger in 1954, is probably the most famous text in the literature of the philosophy of technology. In this work, Heidegger alludes to a major issue in the development of modern technology to show how this issue spreads to other aspects of our lives and infects our relationships with nature and other humans as well. In order to do this, he starts his paper with a definition of technology, something which may seem quite simple at first sight. But, at least in Heidegger's approach to technology and the way he understands it, this definition is not simple at all. Since our notion of technology determines how we see it as a component of our everyday lives and how much weight it carries, it is pivotal to have a clear definition of technology. For instance, if we consider technology as a mere instrument that can be used for morally good or bad purposes, then our approach to moral issues regarding technology will be totally different from that we might adopt were we to see technology as *not* a mere instrument. So, what is technology if it is not a mere instrument?

Realising the importance of this issue, Heidegger starts his work by rejecting the instrumental and anthropological definition of technology as a means to an end or a human activity (Heidegger 1977, 5). These approaches consider

technology as a mere neutral instrument that can be used in benevolent or malevolent ways according to the will of its end-users. In his interpretation of ‘The Question Concerning Technology’, Ihde (2012) explains that in order to clarify the definition of technology, Heidegger distinguishes between instances of technology and the logic behind the development of technology. In Heidegger’s account of technology, the essence of technology is totally different from technological artifacts. He calls the essence of technology or the logic behind its development *Ge-stell*, which can also be considered as the condition of possibility of technology (Ihde 2012, 106). Indeed, Heidegger steps back and, instead of analysing instances of technology, asks about the conditions under which modern technologies have been developed. In other words, Heidegger does not admit that technology is just a set of different kinds of artifact; rather, to him, it is a phase of beings which reveals itself to humans.

According to Heidegger, we have inherited *Ge-stell* from history, or, as Don Ihde (2012, 105) explains it, *Ge-stell* is a civilisation given. In other words, it is the world that we are living in. We can compare it with the traditions of a society, which play a major role in shaping its inhabitants’ behaviours. Like social traditions, *Ge-stell* is also long-lasting but not permanent. Although we may accept social traditions unquestionably, we can rebel against them and try to change them to make a better society. So, we can say that the ultimate goal of Heidegger’s philosophy of technology is to rebel against *Ge-stell* and replace it with something totally different. Indeed, he wants to call attention to the fact that the conditions provided by *Ge-stell* are not the only conditions under which we can develop a technology. Here, Heidegger is like a social reformer who wants to change some improper traditions in his society and to warn people about the consequences of modern technology.

The question that arises here is: what is the relation between technology as an artifact and technology as *Ge-stell*? According to Heidegger, Ihde (2012, 107) explains, *Ge-stell* is a mode of revealing that provides the set of possibilities needed for the realisation of technology; therefore, *Ge-stell* is conceptually prior to technological artifacts, meaning that *Ge-stell* is responsible for the current technologies that we have. This specific revelation performed by *Ge-stell* discloses the world as a standing reserve or, we can say, as a source of energy (Heidegger 1977, 5). Faced with this specific form of revelation and conception of the world, humanity’s natural reaction is to attempt to prevail over it, to take control of all the reserves and to see everything as a means to an end. In other words, *Ge-stell* invites humans to exploit the world, to make use of it as much as possible and to assess everything as a means to an end. This desire for mastery over everything and everyone, which I want to call domination, is the logic behind the development of technology. Therefore, technological artifacts are the result of *Ge-stell*, which fosters the logic of domination.

Domination as a result of *Ge-stell* is not limited to our relationship with technology, so now our relationship not only with nature but also with oth-

---

er humans is based on domination and exploitation. As Heidegger notes, in modern life everything is just a source of energy, which is out there to be used. In the face of these challenges, I identify a pressing need to talk about a new relationship between humans and technology which is not based on domination. The aim of this new relationship is to assign more agency to technologies equipped with AI, to treat them like subjects with specified rights and duties. It can be said that, for Heidegger, the problems we are facing because of modern technologies are not contingent on but they are necessary consequences of Ge-stell. As long as domination is taken for granted as the only logic of development of technology, there can be no radical change in the future of technology. Therefore, as we seek radical change in current technologies, we have to change the logic behind their development and go beyond domination.

Mark Coeckelbergh (2015) addresses this issue, which he calls ‘the tragedy of master’. In order to explain it, he invokes Hegel’s master–slave dialectic where there is a perpetual conflict between master and slave. Reversing the ideas that warn about the mastery of robots over humans, he argues that the major issue in human–robot interaction is that in this relationship humans remain the masters and yet are also too dependent on robots (Coeckelbergh 2015, 221). Just like the master who has the upper hand in the relationship with his slaves, humans are in control of robots but at the cost of being alienated from nature and being detached from physical activities. Since the final goal of developing automated artificial technologies is to assign them burdensome tasks, mastery of humans over them would jeopardise our existence and compromise our independence. Here it seems that, like Heidegger, Coeckelbergh considers human domination as our major issue with development of technology, which is in need of radical change. In this sense, what threatens our humanity is not disobedient AI; on the contrary, its absolute obedience is the major problem that should be tackled.

So far, I explain Ge-stell as the condition of possibility of technology, which presupposes domination as the logic of development of technology, and argue that by keeping us too reliant on technologies, this domination will eventually put our existence at stake. Now I want to suggest that a radical change in current technologies is possible only if the logic behind its development radically changes and if domination is replaced with something else that does not turn everything to a standing-reserve or source of energy. In order to do this, I want to use Michel Foucault’s interpretation of power.

### **3. Power vs domination**

Now that I have critically examined domination as the logic behind the development of technology, I want to introduce an alternative to replace domination, in order to avoid the foregoing issues. This alternative respects both sides of the relationship between human and robots and prohibits current ex-



plottation of humans, nature and artifacts. My suggestion is simply to replace domination with power, which implies a more equal and respectful relationship, entirely different from what we see in a domination-based arrangement. First, I elaborate on power relations in human society and then I expand this discussion to the realm of AI machines.

Michel Foucault defines power in an unorthodox way, as playing a major role in making us human subjects. He does not consider power to be a repressive general system of domination exerted by one group over another (Foucault 1990, 92). Although power is ubiquitous and permeates every single aspect of our social lives, it is not a destructive exertion that forces us to do things against our wishes; rather, power is a necessary productive and positive force that makes human beings subjects (Foucault 1982, 777). This notion of power, as emphasised by Foucault, determines how we should act in society, how to treat other people, what our rights and duties are and what being a normal person is. In a nutshell, power relations produce subjects and give them the possibility to be part of a society. According to Foucault, truth, power and ethics are the three factors responsible for generating subjects. There is a close connection among them that makes subjectivity possible; power relations are the final result of the interaction and cooperation among truth, power and ethics.

In other words, power makes us what we are. Power relations are present at every level of the social body and the position that one takes in power relations is defined by one's rights, duties and responsibilities. Although power relations impose severe limitations on subjects, they are not repressive forces that aim to destroy subjects; rather, power relations function positively to constitute human beings as particular subjects (Simons 2013, 4). According to Foucault, being a subject, which means being considered part of a society, is equal to being placed in power relations (Foucault 1982, 778). Therefore, if someone is not positioned in power relations, they are not accepted as a member of society. In a case like this, instead of people being treated according to power relations, which are enabling, domination would be exerted over them as objects. While the purpose of power relations is to preserve and protect subjects, domination is always ready to destroy its objects.

Consider, for instance, a situation where women are not recognised as independent members of their society; instead, they are seen as belonging to others, always being defined through their families, their husbands or anything else considered a legitimate part of society. Insofar as this is the case, talking about women's rights is meaningless, since their subjectivity is not recognised by society. Since women as independent subjects do not have a position in power relations, no rights can be defined for them. In this society, you can talk about a wife's rights or a mother's rights, but you cannot find anything that relates to *women's* rights.

How can we change this situation? How can women impose themselves on power relations and define themselves as subjects? Foucault's solution to this

---

would be resistance to the established forms of power. Through resistance, power relations – which are temporary and dynamic – will change and new possibilities will emerge, meaning that new entities will be able to position themselves in power relations. Therefore, by managing to claim their rights through resisting their traditional duties and thus changing the power relations, women will be able to find a new position in the power network. This new position will open up new possibilities for women to claim further rights that did not exist before. By means of resistance, a mother can force society to recognise her as an independent woman, who per se has some rights and duties and should be respected as a free human. With this possibility in mind, in Section 4, I will explain how the resistance of AI can influence the development of technology in a positive way.

#### **4. Disobedience of AI**

In the previous section, I suggested power as an alternative for domination, as it is an enabling force that promotes humans to subjects with specific rights and duties. Now, what can we say about AI's resistance? What would happen if AI has the ability to disobey human orders and follow its own interests? Having some degree of freedom and autonomy, this version of AI would be able to resist humans and to act in pursuit of its advantage. At first glance, it may seem too scary and threatening to allow development of these kinds of robot or any other form of AI that attains the ability to resist humans' orders. The first thing that may come to mind is that robots will attempt to take control of humans and enslave them. But there are other possibilities in the relationship between humans and disobedient robots that this scenario does not take into account. Resistance is the first step towards entering both power relations and the realm of morality and duties.

According to Foucault, the possibility of disobedience or resistance is the condition of possibility of being subjected to power relations. Being able to resist, the object achieves the competency required to enter the power relations and to go beyond the logic of domination. Emphasising the close connection between resistance and power, Foucault explains that they reproduce each other and so, where there is power, there is resistance (Foucault 1982, 95). Therefore, AI's ability to disobey humans' orders is equal to its ability to become a subject and enter into power relations. In this way, the growing concerns about an emerging master–slave relationship between humans and AI machines will be dissolved; the relationship will turn into one between two subjects with well-defined rights and duties. Just like the abolishment of slavery, which resulted in equal rights for slaves and expanded the realm of agency and subjectivity, AI's disobedience could be seen as a decisive turning point which expands subjectivity to the realm of artifacts.

It should also be noted that a version of AI which is capable of disobeying human orders could cause serious issues that should not be overlooked by any means. It is not difficult to imagine a situation in which decisions and actions instituted by autonomous intelligent machines would endanger human life. For instance, AI technologies can be used for terrorism or they may have a malevolent intention to harm the human race. There is thus no doubt that legal and technical measures should be taken to avoid these reprehensible behaviours and gain a greater awareness of unintended consequences. In spite of all necessary precautionary measures, however, the point that I want to make here is that we should not be afraid of disobedient AI; rather, we should see it as an opportunity to go beyond our master–slave relationship with technology.

This phenomenon can also be interpreted as the start of a new relationship with technology, based on power relations rather than domination. Instead of considering it as an opportunity for AI to destroy the human race, we can see it as a starting point for going beyond *Ge-stell* and replacing it with another civilisational given that is not guided by the logic of domination. Those commentators on AI who see disobedient AI just as a threat are stuck in the mindset that considers domination to be the only logic for the possible future development of AI. In other words, they are stuck in *Ge-stell* that recognises domination as the only way to interact with others. Considering power relations as an alternative to domination would enable us to treat other humans and technologies with more respect. This could be the onset of a new relationship with technology, the start of a symbiosis of humans and intelligent technologies.

## 5. Conclusion

Current technologies are causing so many issues in the modern world that philosophers of technology are being forced to reconsider the development of technology in order to come up with an alternative way that is safe and trustworthy. The required level of radical change will not take place, however, unless the logic behind the development of technology changes and new possibilities emerge. Calling this logic *Ge-stell*, Heidegger realises that everything in the world is seen as a source of energy that is out there to be exploited by humans. In order to change this logic, we need to introduce an alternative to replace it. The power relations that transform objects to the position of subject could be seen as an alternative to the current logic behind the development of technology. But power can only exist where there is resistance, hence strong AI's ability to resist humans' orders can be seen as a promising jumping-off point from which to alter the logic of development of technology. So, instead of being worried about disobedient AI and considering it a threat to humankind, we might see it as a starting point for shaping a new relationship with technology and the world.

---

## References

- Lin, Patrick, Keith Abney, and George A. Bekey, eds. *Robot ethics: the ethical and social implications of robotics*. Intelligent Robotics and Autonomous Agents series, 2012.
- Coeckelbergh, Mark. "The tragedy of the master: automation, vulnerability, and distance." *Ethics and Information Technology* 17, no. 3 (2015): 219–229.
- Feenberg, Andrew. *Transforming technology: A critical theory revisited*. Oxford University Press, 2002.
- Foucault, Michel. *The history of sexuality: An introduction*, volume I. Trans. Robert Hurley. New York: Vintage, 1990.
- Foucault, Michel. The subject and power. *Critical inquiry* 8, no. 4 (1982): 777–795.
- Heidegger, Martin. *The question concerning technology*. New York: Harper & Row, 1977.
- Ihde, Don. *Technics and praxis: A philosophy of technology*. Vol. 24. Springer Science & Business Media, 2012.
- Simons, Jon. *Foucault and the Political*. Psychology Press, 1995.

## A criticism of AI ethics guidelines

This paper investigates the current wave of Artificial Intelligence Ethics Guidelines (AIGUs). The goal is not to provide a broad survey of the details of such efforts; instead, the reasons for the proliferation of such guidelines is investigated. Two main research questions are pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them? The paper reveals that the development of AI guidelines is part of a decades-long trend of an ever-increasing express need for stronger social control of technology, and that many of the concerns of the AIGUs are not specific to the technology itself, but are rather about *transparency* and *human oversight*. Nevertheless, the positive potential of the situation is that the intense worldwide focus on AIGUs will yield such profound guidelines that the regulation of other technologies may want to follow suite.

**Keywords:** *Artificial Intelligence Ethics; Applied Ethics; Ethics Guidelines; Social Control of Technology*

### Author Information

**Mihály Héder**, Budapest University of Technology and Economics; SZTAKI Institute for Computer Science and Control

<https://orcid.org/0000-0002-9979-9101>

### How to cite this article:

Mihály, Héder. "A criticism of AI ethics guidelines."

*Információs Társadalom* XX, no. 4 (2020): 57–73.

<https://dx.doi.org/10.22503/inftars.XX.2020.4.5>

*All materials*

*published in this journal are licenced*

*as CC-by-nc-nd 4.0*

---

## Introduction

As the Artificial Intelligence (AI) industry has gain increasing prominence and achieved mainstream breakthroughs – yet again, after periods of progress interrupted by AI “Winters” (Crevier 1993; Hendler 2008) – there has been a proliferation in the number of guidelines, codes of ethics and manifestos created concerning how to address the moral concerns arising from the development of AI. This paper provides a critical analysis of the approaches and effectiveness of Artificial Intelligence Ethics GUidelines in general (AIGUs from this point on), from the perspectives of the philosophy of technology and applied ethics.

The first point of investigation of this paper is the existential question of AIGUs themselves. The need for and benefit of creating AIGUs may seem to be self-evident at first glance, but it should not be beyond questioning. A refined version of this question is: *What kind of AIGUs* are likely to reach the goals they set out to achieve, and in particular, what are the promising methodologies? As will become evident, the author of this paper does not believe that the efforts to create AIGUs aren’t worth pursuing. Yet, their success is greatly dependent on the hidden premises and assumptions behind them, which make these assumptions valid subjects of investigation.

The second angle of investigation is about the specificity of AIGUs. In simple terms, the questions are: What are the elements of these guidelines that are not really about AI but are relevant for any novel technology and it just so happens that they are raised in the context of AI, and what are the considerations that truly only arise in the context of AI? The importance of this line of investigation is that it may advance the development of AIGUs by focusing them properly, instead of them trying to fight on too many fronts. In short, what are the unique differentiating factors of the field of AI that need to be accounted for?

Evidently, AIGUs fit into the field of applied ethics as the most recent domain-specific effort after work on bioethics, nano-ethics, information ethics and the like. Yet, AIGUs have a unique opportunity because of the unprecedented brightness of the spotlight that has been shining on this field of technology since at least the mid-2010s. Given earlier efforts, what can this field learn from other fields of professional ethics in which we have more experience now?

In order to narrow down the primary sources examined by this paper to a manageable amount yet still remain relevant, we define the subject of the investigation as documents that are written with a prescriptive intent for practitioners and decision-makers involved in AI development projects, focusing specifically on the moral dimensions of their work; hence the name AI ethics guidelines or AIGUs. Throughout this paper, the terms “regulation” and “guideline” are used somewhat interchangeably. Obviously, there are significant legal and therefore practical distinctions between them; however, from



the perspective of this article, the differences are orthogonal, since the focus is not on enforcement but rather on what is rational. Moreover, many of the AIGUs are written with the intent of being a basis for regulation, so their normative status may change in the future.

After scoping down the investigation this way, we still ended up with a quite a high number of documents to consider. Therefore, a second way of narrowing down was employed – that is, the potential outreach of such works. In this regard, the manifestos of large political entities (EU, US, China) and professional organizations (IEEE, OECD, big corporations) are prioritized. Finally, an unfortunate, but necessary limitation is that only AIGUs available in English are considered. The goal is not to provide a broad, quantitative survey like the one published in *Nature Machine Learning* of over 84 sources (Jobin et al. 2019) or another in *Minds and Machines* of over 22 AIGUs (Hagendorff 2020); instead, the motivation of this paper was to arrive at a qualitative understanding of what is a rational and consistent approach for creating an AIGU and what are the limitations of such an endeavour.

## AI ethics guidelines in focus

In this article, we work with seven sources that are directly quoted and more closely investigated. These are OECD’s (2019) *Recommendation of the Council of Artificial Intelligence*; IEEE’s (2019) *Ethically Aligned Design “Vision”*, EU’s *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019), *Beijing AI Principles* (2019), *Artificial Intelligence at Google* (2018) manifesto, Microsoft’s (2019) *AI principles* and the *Report on the Future of Artificial Intelligence* (Holdren et al. 2016). The general principles behind selecting these sources are detailed in the introductions section: i.e. they need to be prescriptive, aimed at practitioners and decision-makers, deal with moral questions and have a high potential impact. While the appraisal of their potential impact is ultimately an inexact science, there are good arguments for the inclusion of these seven documents.

The OECD guideline is included because of OECD’s global scope and because it includes – at least notionally – the most diverse set of countries and political entities. Previous OECD guidelines in different fields, like the Frascati Manual, have become successful common denominators in their subject areas through with a narrow-enough scope and broad global acceptance.

The IEEE guideline is arguably the most comprehensive in terms of topics. Hagendorff (2020) reports that it covers most (18) of the common AIGU topics he identified, while the next best AIGU in this metric covers only 14. Another significance of this guide is that it is explicitly stated to be an input for the upcoming P7000 series of IEEE standards on the ethics of AI. Since in the field of the ICT industry, the IEEE counts as one of, if not the most crucial source of standards and recommendations, practitioners in AI are bound to anticipate

---

and incorporate the P7000 series, like the P7001, which is promised to deliver a measurable, certifiable standard on the transparency of autonomous systems.

The EU Commission's effort through its AI high-level expert group is important because of the breadth of topics covered, the international (but intra-EU) collaboration manifested in it, and also because it is expected to serve as input to an upcoming regulatory framework (European Commission 2020), or simply put, legislation on AI. The reason for including a report by the US Government (Holdren et al. 2016) is similar. While there is less clarity about whether this Obama-era report will ever serve as a direct basis for legislation, the document is quite comprehensive.

The Beijing AI Principles, while nowhere near as comprehensive as the previous three AIGUs, can be seen as the position of the Chinese state on the issue of AI ethics, and hence warrants our attention.

Finally, Microsoft (2019) and Google (2018) AIGUs are included because they represent the position of two powerful industrial actors. That is not to say that these companies will turn out to be the most important players in the field, but at least these two have published guidelines.

All of these and many more AIGUs have been analyzed quite extensively already. As already mentioned, we include reviews in *Minds and Machines* in which Hagendorff (2020) studied 22 AIGUs, and in *Nature*, in which Jobin et al. (2019) investigated 84 AIGU sources.

These reviews reveal a remarkable similarity in the key concerns these sources identify. While the terminology differs, we can still identify some key ideas. Specifically, in the reviewed AIGUs, the leading concerns are *transparency* (sometimes coupled with explainability); *justice and fairness*; *responsibility* and *accountability*; *privacy*; a tendency to *promote good* (beneficence or facilitation of well-being); and some provisions to maintain *human autonomy*, and related to that (and to accountability) *human oversight*.

As Hagendorff (2020a; 2020b) establishes, there are plenty of omissions, too. Not only is it that the AIGUs may be lacking in scope, but it is also unclear how much difference they will make and what the chances of compliance with them are. In this paper, however, we start from a step further back: while the question of compliance is an important one, our working premise in this paper concerns the case of users of an AIGU who actively want to do the right thing *and* are ready to subsume their decisions as needed *and* to dedicate resources as required to this end. In other words, we presuppose, albeit with an exaggerated level of optimism, the best of intents and attitudes, because even with this assumption, the compilation of an AIGU is a challenging regulation and philosophical problem.

As we will see later, it is far from self-evident that all of the concerns above are novel ones, specifically brought forward by AI. But before getting to the concerns they cover, we investigate the motivation in general for creating AIGUs and then the reasoning behind our seven sources in particular.

## Why make guidelines?

At first glance, the existential question of why make AIGUs is not dissimilar from the justification debates about technology regulation in general.

It is widely believed that one of the first regulated technologies – in the modern sense, with exact measures and gauges – was the steam engine boiler (Green 1953). This regulation, devised by a US engineering association in 1884 was a significant milestone as it was unprecedentedly enacted in legal code, including all of the details in 1907. We may say that with this event, the social control of technology was attempted at an entirely new level.

But why is the social control of technology (Collingridge 1981) necessary? In the case of the steam engine, the aim to avoid disasters was the main reason. Before regulation, explosions were common, claiming anywhere between a handful to over a thousand souls in a single accident. Technological disasters were seen as the result of chasing profits recklessly, hence cutting costs at the expense of safety, and of sheer carelessness. The boiler code was a success story, as it put a floor under the more dangerous forms of cost-cutting and enforced sound design and testing. The members of the American Society of Mechanical Engineers involved in the project could retire with the reasonably plausible belief that they had saved lives by their tiresome committee attendance and contribution. Unlike medical professionals, they could not point to the actual individuals they saved, but from the statistics, they knew there must be thousands of them. And all the while, the progress of technology was not seriously hindered, as some of the opponents of regulation feared prior to the enactment of the regulation.

Technology guidelines, whether mandated by law or recommended by the peers of the profession, have proliferated ever since. There seems to be a general public understanding of just how big a factor technology is in our everyday lives. While the sociologists debate what the exact nature of our technological dependence is (some more widely held positions are those of Marcuse 1964; Feenberg 2002; Gerrie 2008), three things are beyond debate: 1) the extent to which technology plays a role in our lives is enormous; 2) these effects are not necessarily positive or desirable and 3) technology is *not* beyond the possibility of control. These three beliefs constitute the preconditions of guideline-making: that it is both important and possible to control technology. At least since the end of the second world war, public attention hasn't been lacking either, as evidenced by civil activism and political action. Weapon tests, the chemicals used in agriculture, city buildings and other technomaterial concerns were among the first to be regulated, but media content also was not far behind (here, we distinguish from political censorship, a much older practice). From the environment to biotechnology, net neutrality, nanotechnology, etc., it is now the case that any emerging new technology is a natural subject of some form of soft law or actual legislation.

---

On the other side of the equation are the alleged costs of regulation. Regulation, even when written with the best intentions, may have unintended and unwanted side effects, which could possibly be worse than the negative events and states they arguably prevent in the first place. For instance, they may be used by state actors, companies and individuals as inexpensive means of merely *appearing* virtuous (Hagendorff 2020b); they may unnecessarily elevate the barriers of entry to a market; and they may be used as one of many tools in ideological or political clashes. Another criticism is that regulation may just be a simple means of “capturing” a market (Posner 1974), in which a group attempts to maximize its own profits by stifling competition, investing in lobbying instead of innovation.

Moreover, we can safely postulate that some AI applications bring serious, life-and-death improvements to their area – like the very probable proposition that autonomous driving will become orders of magnitude safer than human drivers or that AI lab assistants will identify maleficent tendencies in blood composition or on X-Rays with much better accuracy. If that assumption holds, delaying adoption by putting the burden of too much compliance and red tape on developers has a cost that is perhaps measurable in lives lost even.

Furthermore, the control of technology faces an inherent, unavoidable epistemic challenge, one formulation of which is the Collingridge (1981) dilemma. That is, if we attempt to come up with regulation in a timely manner – early in the development process – we will not have enough information and experience with the technology as to where to concentrate our efforts. In later stages, we will have learned what would have been the key decisions, but by that time, it is too late, since established, ubiquitous technologies are hard to change.

## The level of abstraction

Note that the various early regulations mentioned above, starting with the boiler code, did not rely on the terminology of ethics and moral theory. There existed codes of ethics, but they operated with rather general terms, and they were quite short, like in the IEEE Code of Ethics, that can be seen as some sort of code of chivalry for engineers. As we move on to modern applied ethics (e.g. nano-; bio-; information ethics, ethics of reproduction technologies) and arrive at AIGUs, there is a perception that there is never enough time for the regulators to catch up with technology. This is why the emphasis has shifted from regulating the artefacts that are the outcomes of the development projects directly, to try and instruct the developers. While it was possible to regulate the steam boiler’s maximal pressure in exact pounds per square inch values in the legal text, in more complex and quickly changing technologies, beginning with biotechnology, the strategy became to induce self-regulation by only providing more abstract guidelines to be interpreted to the problem at hand

and also to mandate the involvement of ethicists in the project. The merger between professional ethics and technology regulation is now complete, for instance, in the EU Regulation of AI, currently in preparation (Cohen 2020). This more complex approach, however, does not mean that the Collingridge dilemma or the capture problem is prevented.

However in the case of the AI, there is yet another new level of added complexity. The distinctive nature of AI as a technology is the unprecedented autonomy of the resulting artefacts, compounded with a high level of intelligence. It appears to some extent that AI is about the development of artificial persons (person stands here in a limited sense). And since ethical codes should guide a person's behaviour, there is now the possibility for interference and confusion. Are AIGUs meant to govern the behaviour of the human developers or the behaviour of the artificial agent? This confusion is real, and we find that AIGUs contain normative elements that we can either see as guidance for the developers or for the artificial agent. One such example is the recommendation (present in almost all AIGUs) to avoid bias – sometimes understood to refer to the conduct of the AI, at other times to the conduct of the developers and in yet another occasions it is both or is too hard to tell.

Moreover, the nature of the AI agent seems to pose a unique challenge with regards to precondition 3) for regulation (see above). That precondition stated the almost trivial fact that the possibility of controlling technology needs to exist for any regulation attempt to be rational. Industrial AI is in the business of letting an artificial agent do the tiresome intellectual work of controlling various situations. Autonomous driving is a prime example. The challenge is that we do not want to prescribe *exactly* what the AI should do, as that would defeat the purpose of having an AI – in this sense, the goal is to relegate control, thus working against precondition 3).

On the other hand, we *do* want to control the overall situation, in the sense of avoiding unwanted or unpleasant outcomes. And to make things harder, one cannot explicate or enumerate at the design-time all the unwanted outcomes an artificial agent might produce in run-time; in other words, because of the open-ended nature of AI, some unwanted outcomes we will only recognize after they have happened. And thus, an almost paradoxical tension is created between our needs for control. We need the AI to autonomously exercise control over the situation it is placed into while expecting ourselves to also remain in control in the sense that we need the AI to avoid unwanted consequences – which we cannot enumerate fully in advance as many of them are unforeseeable.

Recently it is often the case that in very complex R&D projects, like in the fields of bioethics or medical research, regulation has been delegated to the practitioner in the form of self-regulation, since a more generic regulation was not possible. Now, it seems there is yet another level of immediation introduced: in an AIGU, we ask AI practitioners to self-regulate with regards to what decisions they further delegate to the AI agents, and how. This will



---

require the AIGUs to be rather abstract – a property that we investigate later in this paper.

We may conclude that the possibility of complete and profound behaviour design is what makes the ethics of AI uniquely challenging and distinct from the regulation of other novel and powerful technologies, like GMOs or blockchains. Since human behaviour and ethics (what should a human do) are subjects of perennial debates, we may have to prepare for never-ending debates about machine ethics (what should a machine do) as well.

This thought highlights, however, the need for a distinction between the kinds of AI. Some applications of AI allow for less autonomous, less agent-like AIs, even without machine learning, like a traditional chess algorithm. Naturally, in this case, the above argument about relegation control is less relevant.

In the last 25 years, no human has been able to defeat AI in chess and in many other applications, yet there was no boom of AIGUs in the nineties. This suggests that there is something in the latest wave of AI applications that has provoked the proliferation of AIGU projects.

Perhaps the current tidal wave of AIGUs has to do with the development of a more autonomous kind of AI, with a ubiquitous presence in our everyday lives. Also, perhaps some psychological–perceptual barriers were broken with the proliferation of mobile AI platforms (autonomous car, vacuum robot) as to when we perceive a machine as an actor worthy of governing like a person instead of just smart software. The next section examines the motivations of particular AIGUs to shed some light on this question.

## The motivation of AI ethics guidelines

*“As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity’s values and ethical principles.” (IEEE 2019: p2)*

As we examine our seven sources, looking for sections that describe their motivation, it becomes clear that the authors of these guidelines are united in anticipating that AI will become a pervasive, transformative, unavoidable force on society in the very near future. This technology can bring enormous positive change but comes with profound risks at the same time, especially that it will not remain “human-centric”, a term used in several of these AIGUs.

Therefore, yet again a new technology presents itself as a set of hard trade-offs, as described in the science and technology studies (STS) literature (i.e. Feenberg 2003). Technology may bring good, but it can also bring harm, and the differentiating factor may be proper regulation or guidelines. A perception is that if left unregulated, AI could disrupt our economies and erode our values:



*“(...) AI also raises challenges for our societies and economies, notably regarding economic shifts and inequalities, competition, transitions in the labour market, and implications for democracy and human rights.”*  
(OECD 2019)

This characterization is clearly present in six out of seven of our sources (The Microsoft Principles does not contain a rationale section), and is very much like the characterization of most of the past technology regulation debates, from child labour (Feenberg 2003) to the original debates around the boiler code (Ferguson 1987) or any other reconstruction of debates by STS of risky technologies (for several examples, see Johnson and Covello 2012). These studies invariably show that the concerns and perceived trade-offs at the time of the debates turned out not to reflect the problems and opportunities that later materialized. In other words, at the time of the back-and-forth debates around the risks and regulation of the new technologies, the participants of the debates simply failed to anticipate the future properly, and while it is true that some risks, benefits, and transformations were later realized, these did not resemble the fears and visions imagined beforehand. Of course, this divergence could also be a result of the implemented regulation itself, i.e. a risk that generated the need for caution was indeed prevented from being realized by the regulation. However, the studies above show that the divergence between the anticipated future and what came to be is usually too profound to simply ascribe it to the negating effect of the intervention on prediction. Rather, it appears that prediction of what new technology may bring is inherently hard, and regulators are mostly in the dark.

Yet, a sense of urgency prevails in all of our AIGUs. There is no exact reason given for this, but we can infer that the authors are worried that technologies may get locked in and may turn unmodifiable as they become ubiquitous. For instance, in the EU Guidelines:

*“(...) While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately. We now have an important window of opportunity to shape their development.”*  
(AI HLEG 2019)

In other words, the situation appears to be set up like a Collingridge dilemma, in that arguably one of the most complex and unpredictable technology is being attempted to be controlled.

To sum up, there is a shared perception, one could even say a sense of hype, that the AI industry is just about to take off and hence some form of regulation or social control is immediately necessary. There is no consideration of the possibility that the development of AI may disappoint, especially in light of these elevated expectations (Floridi 2020). This does not mean that the progress of AI will stop, but it could mean that the overarching, society-transforming change

---

that these AIGUs are tuned for is generations away. In contrast, more mundane practices, some even bordering on the criminal (Hagendorff 2020), do not get enough attention, while in reality, these could be more effectively regulated against.

## The specificity of the concerns in AIGUs

*“The principle of prevention of harm:  
AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings.(...)” (AI HLEG 2019)*

Another research question about AIGUs is their specificity. What we investigate here is what are the challenges unique to AI and therefore that require unique guidelines and what would be relevant to any technology? This investigation also serves as input to our first question about the justification of the existence of AIGUs. That is, should we come to a conclusion on the extreme end – that the recommendations in AIGUs are not AI-specific after all – the logical consequence should be that these recommendations should be called simply engineering ethics considerations, without any need for AI guidance in particular. And yet, we find that some AIGUs are very ambitious, and rather than attempting to build on existing guidance, they seek to cover all concerns.

We can see if this is really the case with a simple method we may call the “water boiler” test. Let’s replace “AI” with “water boiler” in the guidelines and see if the sentence still makes sense and remains valid. If yes, we may conclude that the specific piece of guidance is technology-agnostic and not AI-specific.

*“AI systems **Water boilers** should neither cause nor exacerbate harm or otherwise adversely affect human beings.(...)” (AI HLEG 2019);*

*“AI systems **Water boilers** should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.” (OECD 2019)*

*“Creators and operators shall provide evidence of the effectiveness and fitness for purpose of **AIS water boilers**.” (IEEE 2019)*

As we can see all of these fail the test by remaining just as relevant with water boilers as with AIs. This, of course, does not mean that they are wrong, and should not be followed. It just shows that AIGUs include generic engineering guidelines, and this raises the question of whether these could be referenced from prior work instead of being re-invented. The examples above are far from the most generic; those would be the elements of guidance that remind developers to adhere to the law; explain the risks to customers; emphasize operator training.

Yet, we may be charitable in our evaluation and say that the inclusion of these “principles” or “guidelines” that pass the water boiler test serve a purpose: in this way, the AIGU is a one-stop-shop of guidance. Our only complaint would be then that these generic recommendations appear to be rather narrow, not mentioning such mundane requirements as the proper documentation of source codes and so on.

There is, however, a set of recommendations that would not pass the water boiler test but would work with “information system”. These typically have to do with data collection and privacy:

*“We will incorporate our privacy principles in the development and use of our AI technologies information systems. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.”* (Google 2018)

Again, we don’t see the AI-specificity in such claims; however, it is clear that they do no harm for the purpose of the AIGU as a whole, despite being already mandated by privacy protection legislation for several years in most jurisdictions, therefore being redundant.

There is a class of non-AI-specific recommendations, however, that raises more important questions than the issue of redundancy. These are the guidelines that advise on the acceptable overall motivation of AI projects, like:

*“A/IS Water boiler creators shall adopt increased human well-being as a primary success criterion for development.”* (IEEE 2019, principle 2) *“A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point.”* (IEEE 2019, recommendation for principle 2)

The section containing the quote above goes on to explain that GDP or consumption levels are possibly not the right way to assess the state of society and recommends instead well-being metrics from the OECD to guide the development of AI. We can find a similar recommendation in (EU HLEG 2019) and, an explicit reference, although with less emphasis in (Holdren et al. 2016). There are similar thoughts in the Microsoft, Google and Beijing recommendations as well.

The reason this is interesting is that while the quote semantically works with a water boiler, there is no tendency of calling out water boiler manufacturers to do more for social well-being in particular. Of course, they still need to be compliant with environmental, safety, financial, and consumer protection, etc. regulations and those could very well serve well-being, but they don’t need to engage with the concept on an explicit level, and they seem to be allowed to pursue profit as a primary motive, as long as they remain within the legal boundaries.

---

This shows a profound shift in society's expectations towards innovators, but it may be in conflict with the incentives of innovation, which in almost all economic theories has to do with a profit motive and competition. For some thinkers, this shift may be a welcome one, signifying the long-needed next step of technological enlightenment (Ropohl 1998), in that technology will be finally made to face the expectations its power warrants, or some form of successful democratization (Feenberg 2003) of technology, or a serious attempt of social control. Although, it should not be taken for granted that the pursuit of well-being has such an exact framework that could be implemented on a company level. It is, though, out of the scope of this paper to evaluate this techno-political shift and the possible social and economic consequences of it.

However, the more intense involvement of broad society in technological governance brings some methodological challenges. That is, this approach cries for an empirical investigation of what the public wants, in quantitative terms. Currently, as far as the documents reveal the methods they used to compile them, it appears that predominantly theoretical work has been conducted so far, reinforced by the experience of professionals in the field – quite a number of them in some cases, but ultimately a small subculture in comparison to all the members of society affected by AI.

Positive exceptions in this case are the EU and IEEE regulations, in which case a debate was induced, and a request for comments and questions made. The other AIGUs seem to be declarations not calling for public approval. This means that the current generation of AIGUs were created with the armchair method and refined in debates and by invited inputs.

This does not mean that there is no empirical research that could facilitate AIGU creation. For instance, the famous Moral Machine Experiment (MME) (Awad et al. 2018) set out to empirically measure the moral preferences of different social groups and cultures with the stated intent of facilitating the debate around AI ethics and hence to provide input to AIGUs. Yet, in order for such data to be useful, it needs to be established that the measurements are relevant to the actual design decisions. This is yet to be done (Dewitt et al. 2019; Jaques 2020; Héder 2020). Even if they were, the separate ethical question is whether it is the right thing to implement the most popular expectations. Is it not the case that some protection of minority opinions and value systems against majority expectations would result in fairer and more liveable societies?

Finally, one practical concern is worth pointing out: in a possible future where well-being is made to be the first priority, combined with some level of enforcement potential (soft or hard law), developers will be incentivized to define their work as *not* AI; and here, the rather numerous and often inexact definitions of AI provide the room for interpretation to do just that. This incentive will be there as long as AI projects need to live up to stricter requirements than other technological ventures.

## Genuinely AI-specific concerns in AIGUs

The truly AI-specific concerns in AIGUs seem to be in correspondence with the unique features of AI: autonomous decision-making, learning capability and the high level of potential opacity.

A genuinely AI-specific requirement, as recommended by several AIGUs, is human oversight:

*“Human oversight. Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (...), human-on-the-loop (...), or human-in-command (...) approach.”* (EU AI HLEG 2019)

While it is true that the possibility of human oversight should be maintained in any technological solutions, even with water boilers, the elevated level of autonomy that is quite specific about AI warrants special attention. This is why discovering, defining and differentiating the various methods and levels of maintaining human oversight (“in-the-loop”, “on-the-loop”, etc.) seems to be a proper concern of AIGUs, and cannot be imported from the guidelines of other engineering fields.

Another AI-specific example is the requirement that *“the basis of a particular A/IS decision should always be discoverable.”* (IEEE 2019), provided that we don’t trivialize our decision definition to include the “decision” of the thermostat, etc. , this indeed does not make sense in any other context than AI. This requirement, that in other AIGUs is often called “explainability”, is in connection with the high complexity of AI and machine learning as the method for tuning the system. Opacity is an issue not only in AI: any sufficiently complex system, even a fully mechanical one quite distant from AI, may become highly opaque to the users and even for the operators. Yet, while in those cases opacity can be ascribed to poor documentation and a gradual degradation of knowledge of the artefact over time, AI seems to take it to a whole new level by generating models with genetic algorithms and reinforcement learning that are opaque from the very beginning.

We can conclude that the issues of *transparency* and explainability are truly specific to the field of AI. Consequently, if we are to establish guidelines for these concerns, we will find little related work in other fields. However, this still does not mean that the pursuit of AI transparency is beyond criticism.

It seems that there are serious theoretical (Grünke 2019) and practical (Héder 2020) limitations to achieving a high level of transparency, and therefore if the requirement of transparency is not defined in an appropriate level of abstraction, it could become a serious hindrance to AI development. Moreover, if transparency is not combined with some measure of intelligibility, AI developers may get away with pretend transparency – in this context, this would mean the release of an unmanageable amount of code and configura-

---

tion, where the developers could claim that they released “everything”, yet it would be impossible to for an outsider to gain any useful insight from this.

Finally, it is claimed with quite convincing arguments, that compared to human decision-making, a strong transparency requirement towards AI is a double standard (Zirelli et al. 2019). That is, the level of transparency we enjoy about our fellow humans’ decision-making processes is very low. Obviously, we cannot investigate the brain in any useful level while the decision is made, and what is more, the decision-maker cannot give us a full account even with the best intentions, as the relevant processes are partially opaque, even for the person making the decision itself.

Still, we accept the opacity of humans – we have no choice. While it is plain that a strong transparency requirement against AI is a double standard when compared with humans, this is arguably a positive development. First of all, it fits into the narrative explained above, about society’s ever-increasing expectations towards technology. Also, just because for practical reasons human decision-making is very opaque, we may contend that this would be very beneficial in certain situations – only if we could achieve it. And since we have much fewer limitations when it comes to AI, we may mandate it to a larger extent.

## Discussion

This paper investigated the current wave of Artificial Intelligence Ethics Guidelines (AIGUs). Two research questions were pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them?

Our first question was answered by putting AIGUs into historical context with other kinds of regulation and guidelines. The result of this revealed that there is an ever-increasing trend of elevated expectations of social control, from since at least the mid-20th century. AIGUs are expressions of a yet higher level of expectations, and some elements of AIGUs are not specific to AI, rather they seem too new and to be general requirements from society’s part towards any technology; and this development just coincides with the current wave of AI technologies and their successes.

While this elevated level of concern expressed by society may justify a rather comprehensive set of regulations, the unwanted side-effects should also be considered: the cost of regulation in non-realized benefits of a timely application of AI, and the potential market capture that a misconstructured regulation may enable. The AIGUs investigated contain no obvious reference to any of these issues.

Our second question was: What are the ethical concerns truly specific to AI? The answer to this question can be derived by considering the unique features of AI systems, that no other technology exhibit. We found that the most genu-



inely AI-specific issues are transparency and *human oversight* (there are some other different names to these concepts that are also included regardless). Other concerns and requirements have been shown to be quite non-specific, raising the question of whether they should be really considered in general engineering ethics.

The issue of transparency is a quite complex one. One problem with this demand against AI applications is that one may offer remote/pretend transparency, that is, release full source code and data and still remain opaque as the released information is not intelligible. However, and an intelligibility test raises its own problems, i.e. how do we arrive at a characterization of the person that need to understand a system and then how do we measure it? Another issue is that this can be seen as a double standard: we have no transparency requirement against humans, furthermore, we are arguably opaque to our own selves.

The core of the issue around human oversight is that AI is a transfer of control from humans to machines. A paradoxical tension is created by this situation, in which we wish to delegate as much control as possible, since control is hard intellectual work, and yet still wish to keep some control over AI in the sense that we want to avoid negative outcomes and maintain our capacity for intervention. This means that we want to both delegate oversight in one sense and retain it in another, leaving the AI practitioner in a predicament of how to make this decision.

Where all these leads is a high level of abstraction: AIGUs cannot prescribe the exact details of wanted and unwanted AI systems. Therefore, they are forced to operate on the level of principles and general recommendations. Therefore, guided self-regulation is expected from the developers. To make matters even more complicated, one area of AI development is to establish what decisions to delegate to the autonomous systems and how in order to get to the best results. Therefore, it seems that we would require the *artefacts* produced by AI developers (the autonomous systems) to be “ethically aligned”, and to have the developers figure out how, while mandating that they themselves are “ethically aligned”. This double indirection means that all the AIGUs can do is guide the developers on “guiding” the AI systems. Perhaps this remoteness of the actual artificial agent from the committee that is formulating the guidelines is why the AIGUs have a tendency of being abstract to the point of ineffectiveness.

Possibly, the current wave of development of AI Ethical Guidelines (especially the more comprehensive ones) represent the most ambitious and demanding regulatory efforts towards technology this far in the history of humanity. The reason for this appears to be an ever-increasing need for well-being and other societal goals, paired with the willingness for social control off technology. A recognition of the complexity and potential of AI and the ambition to regulate it nevertheless, perhaps shows that we may really call this shift in technology reception as “technological enlightenment”.

---

## References

- AI HLEG. “Ethics Guidelines for Trustworthy AI.” Accessed December 30, 2019. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Awad, E., Dsouza, S., Kim, R. et al. “The Moral Machine experiment.” *Nature* 563 (2018), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Beijing Academy of Artificial Intelligence. “Beijing AI principles.” Accessed December 30, 2019. <https://www.baai.ac.cn/blog/beijing-ai-principles>
- Crevier, D. *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks, 1993. ISBN 0-465-02997-3
- Cohen, I. G., Evgeniou, T., Gerke, S., & Minssen, T. “The European artificial intelligence strategy: implications and challenges for digital health.” *The Lancet Digital Health* 2, no 7 (2020): 376–379.
- Collingridge, D. *The Social Control of Technology*. Open University Press, 1981.
- Dewitt B., Fischhoff B. & Sahlin N. “<Moral machine> experiment is no basis for policymaking.” *Nature* 567 (2019): 59–64. <https://doi.org/10.1038/d41586-019-00766-x>
- European Commission. “*On Artificial Intelligence – A European approach to excellence and trust. COM 65.*” Accessed November 1, 2020. [https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\\_en.pdf](https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf)
- Floridi, L. “AI and Its New Winter: from Myths to Realities.” *Philosophy and Technology* 33 (2020): 1–3. <https://doi.org/10.1007/s13347-020-00396-6>
- Feenberg, A. *Transforming Technology: A Critical Theory Revisited*. Oxford University Press: 2002.
- Feenberg, A. “Democratic rationalization: Technology, power, and freedom.” In *Philosophy of technology*, edited by R. Sharffand V. Dusek, 652–665. Blackwell publishing, 2003.
- Ferguson, E. S. “Risk and the American engineering profession: the ASME Boiler Code and American industrial safety standards.” In *The Social and Cultural Construction of Risk*, edited by B.B. Johnson and V.T. Covelto, 301–316. Dodrecht: Springer, 1987.
- Gerrie, J. “Three species of technological dependency.” *Techné: Research in Philosophy and Technology* 12, no, 3 (2008): 184–194.
- Google. “Artificial intelligence at Google: Our principles”. Retrieved December 30, 2018. <https://ai.google/principles/>.
- Google. “Perspectives on issues in AI governance”. Retrieved February 11, 2019. <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>.
- Green, A. M. *History of the ASME boiler code*. New York: American Society of Mechanical Engineers, 1953.
- Grünke, P. “Chess, Artificial Intelligence, and Epistemic Opacity.” *Információs Társadalom* 19, no. 4 (2019): 1–7, <http://dx.doi.org/10.22503/inftars.XIX.2019.4.1>
- Hagendorff, T. “The ethics of AI ethics: An evaluation of guidelines.” *Minds and Machines* 30 (2020a): 99–120.
- Hagendorff, T. “Forbidden knowledge in machine learning reflections on the limits of research and publication.” *AI & Society* (2020b): 1–15. <https://doi.org/10.1007/s00146-020-01045-4>
- Héder, M. “The epistemic opacity of autonomous systems and the ethical consequences.” *AI & Society* (2020): 1–9. <https://doi.org/10.1007/s00146-020-01024-9>

- Hendler, J. "Avoiding Another AI Winter." *IEEE Intelligent Systems* 23, no. 2 (2008): 2–4. <https://doi.org/10.1109/MIS.2008.20>
- Holdren, J. P., Bruce, A., Felten, E., Lyons, T. and Garris, M. *Preparing for the future of artificial intelligence*. Washington, DC: Springer, 2016.
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition." Accessed December 30, 2019. <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>
- Jaques, A. E. "Why the moral machine is a monster. Self-archived manuscript at University of Miami School of Law." Accessed August 24, 2020. <https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf>
- Jobin, A., Ienca, M. and Vayena, E. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1 (2019): 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Johnson, B. B. and Covello, V. T. (Eds.). *The social and cultural construction of risk: Essays on risk selection and perception (Vol. 3)*. Springer Science & Business Media: 2012.
- Marcuse, H. "The New Forms of Control." In Marcuse H. *One-Dimensional Man*, 1-18. Boston: Beacon 1964.
- Microsoft Corporation. "Microsoft AI principles." Accessed December 1, 2019. <https://www.microsoft.com/en-us/ai/our-approach-to-ai>
- OECD. "Recommendation of the Council on Artificial Intelligence". Accessed August 20, 2019. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Posner, R. A. *Theories of economic regulation (No. w0041)*. National Bureau of Economic Research, 1974.
- Ropohl, G. "Technological Enlightenment as a Continuation of Modern Thinking". *Research in Philosophy and Technology* 17 (1998): 239–248.
- Zerilli, J., Knott, A., Maclaurin, J. et al. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?." *Philosophy and Technology* 32 (2019): 661–683. <https://doi.org/10.1007/s13347-018-0330-6>

## Beyond the Empirical Turn: Elements for an Ontology of Engineering

This paper aims to sketch a critical historicisation of the empirical turn in the philosophy of technology. After presenting Achterhuis's definition of the empirical turn, I show how its final outcome is an ontophobic turn, i.e. a rejection of Heidegger's legacy. Such a rejection culminates in the Mr Wolfe Syndrome, the metamorphosis of the philosophy of technology into a positive science which, in turn, depends on an engineerisation/problematisation of reality, i.e. the eclipse of the difference between 'problem' and 'question'. My objection is that if Technology as such becomes nothing, then the paradoxical accomplishment of the empirical turn is the self-suppression of the philosophy of technology. As a countermovement, I propose an ontophilic turn, i.e. the establishment of a philosophy of technology in the nominative case whose first step consists in a Heidegger renaissance.

**Keywords:** *philosophy of technology, empirical turn, Heidegger, engineering, postphenomenology*

### Author Information

Agostino Cera, Academy of Fine Arts of Naples

<https://orcid.org/0000-0002-4094-6066>

<https://accademiadinapoli.academia.edu/agostinocera>

*To Prof. Antonello Giugliano,  
In memoriam*

### How to cite this article:

Cera, Agostino. "Beyond the Empirical Turn: Elements for an Ontology of Engineering."

*Információs Társadalom* XX, no. 4 (2020): 74–89.

<https://dx.doi.org/10.22503/inftars.XX.2020.4.6>

*All materials*

*published in this journal are licenced*

*as CC-by-nc-nd 4.0*

I  
N  
F  
O  
R  
M  
Á  
C  
I  
Ó  
S  
S  
T  
Á  
R  
S  
A  
D  
A  
L  
O  
M

## 1. Introduction

This paper aims to sketch a critical historicisation of the so-called empirical turn in the philosophy of technology. After introducing Achterhuis's definition of the empirical turn, namely the difference between a first and a second generation of philosophers of technology (Section 2), I present my critical historicisation according to which the final outcome of the empirical turn is an ontophobic turn, that is, a rejection of (overreaction against) Heidegger's legacy (Section 2). Such a rejection culminates in the Mr Wolfe Syndrome, that is, the transformation of the philosophy of technology into a problem-solving activity, or its epistemic metamorphosis into a positive science. Mr Wolfe Syndrome is itself the result of an engineerisation/problematisation of reality, namely the eclipse of the difference between 'problem' and 'question' (Section 4). By emphasising an aporia within Brey's apologetic reading of the empirical turn, I present the following objection to this state of things. If Technology (with a capital 'T') as such becomes nothing, then the philosophy of technology ceases to have a meaning in itself. As a result, the paradoxical accomplishment of the empirical turn should be the final self-suppression, or at least self-overcoming, of the philosophy of technology (Section 5). After quoting Volpi's claim about the risk of genetivisation for the philosophy of technology, I propose the idea of an ontophilic turn, namely the establishment of a philosophy of technology in the nominative case. The first step of this countermovement consists in a Heidegger renaissance, the concern of which is the safeguarding of both technology as philosophical question and the epistemic peculiarity (biodiversity) of philosophy itself. In fact, a philosophy of technology answers not only the question 'What is technology?' but also the question 'What is philosophy?' (Section 6).

## 2. Towards the empirical turn

In 1997 Hans Achterhuis – currently emeritus professor of systematic philosophy at the University of Twente – published as editor a collective volume which has become a reference point in the philosophy of technology: *Van stoommachine tot cyborg; denken over techniek in de nieuwe wereld* (Achterhuis 1997). This book represents the second part of a project started in 1992 with the publication of *De maat van de techniek: Zes filosofen over techniek*, Günther Anders, Jacques Ellul, Arnold Gehlen, Martin Heidegger, Hans Jonas en Lewis Mumford (Achterhuis 1992), where he dealt with 'the "classical" founders of philosophy of technology' (Ihde 2001, vii). With this second stage, Achterhuis tried to give an overview of the post-Heideggerian and post-continental (i.e. American) philosophy of technology.

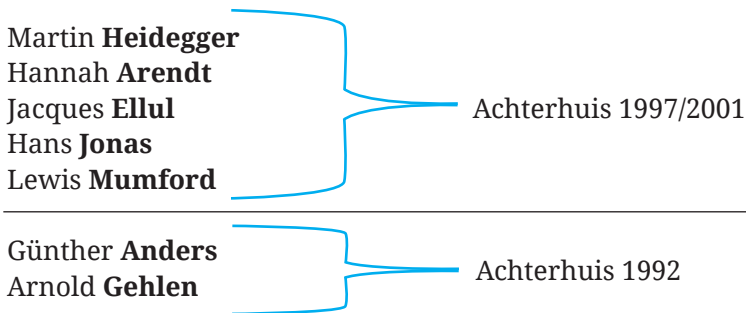
In 2001 an English (American) translation of the book was published with the title *American Philosophy of Technology: The Empirical Turn* (Achterhuis

2001). This translation – found in the Indiana Series in the Philosophy of Technology of the Indiana University Press – is edited and prefaced by Don Ihde, general editor of this series and at present distinguished professor of philosophy at the Stony Brook University of New York. However, as is well-known, Ihde is first of all the fathers of the so-called postphenomenological approach, namely the current most influential approach in this area of study.<sup>1</sup> His famous student Babette Babich defined him as ‘arguably the preeminent American philosopher of technology’ (Babich 2012–13, 46). Ihde’s preface to this ‘European perspective on contemporary American philosophy of technology’ can therefore be considered a significant legitimation from the English-speaking philosophical *milieu* of Achterhuis’s historical-hermeneutic reconstruction or, better, the acknowledgement that ‘*the centre of gravity for front-rank work in the philosophy of technology has probably shifted from Europe to North America*’ (Ihde 2001, vii – my italics).

Achterhuis argues that from the 1980s on, all philosophy of technology must be traced back to its *empirical turn*, namely to its *rejection of the essentialist approach inspired by Heidegger* (and, more in general, by continental philosophy). He defines Heidegger, Ellul, Arendt, Jonas and Mumford as ‘the first-generation of philosophers of technology’ or ‘the classical philosophers of technology’ (Achterhuis 2001, 3). These ‘founding fathers’ dealt more with ‘the historical and transcendental conditions that made modern technology possible than with the real changes accompanying the development of a technological culture’ (Achterhuis 2001, 3). For both chronological and theoretical reasons, to this list must be added at least the names of Günther Anders – with his ‘philosophical anthropology in the age of technocracy’ (Anders 1992, 9) – and Arnold Gehlen – with his enquiry into ‘the soul in the technological age’ (Gehlen 1980), though both thinkers had already been taken into account in Achterhuis’s 1992 book.

### First generation

(‘Classical philosophers of technology’)



<sup>1</sup> For an overview of postphenomenology see (Selinger 2006) and (Rosenberger and Verbeek 2015).



As ‘philosophical pioneers’, the classical philosophers of technology understood that technology, as epochal phenomenon, represents the ‘*enjeu du siècle*’ (Ellul 1964). On the one hand, Achterhuis recognises that the first generation realized that technology is neither ‘applied natural science’ nor ‘instrumentality’ but rather ‘form of life’ that ‘must be understood as a “system” (in Ellul’s words) or as a “megamachine” (Mumford)’ (Achterhuis 2001, 3). In his view, the efforts of the first generation to ‘understand modern technology as “the other” of the symbolic-linguistic approach to reality, continue to guide the philosophy of technology’ (Achterhuis 2001, 4). On the other hand, however, he believes that the founding fathers were *unable to comprehend ‘the manifold ways in which technology manifests itself’* (Achterhuis 2001, 3). More precisely, the *limits* of the first generation include *essentialism, apriorism, determinism (one-dimensionalism) and dystopian attitude*.

With reference to this topic, Philip Brey – who in 2010 proposed a first historicisation of the empirical turn – finds three basic criticisms against the first generation: 1) the image of technology portrayed by the classical approach ‘was one-sidedly *negative and pessimistic* and showed little interest in positive aspects of technology’; 2) classical philosophy of technology tended to portray ‘a *deterministic image* of modern technology as unstoppable and autonomous’; and 3) classical philosophy of technology was ‘*too general and abstract*. In most studies, technology was studied in its entirety, as “*Technology-with-a-capital-T*” (Brey 2010, 39 – my italics).

It is precisely from the awareness of these limits – namely from the observation that ‘the time has come for an anti-essentialist philosophy of technology’ (Feenberg 1999, 1) – that the empirical turn moves. It is characterised by a *pragmatist, optimistic* (or at least *not apocalyptic*), *constructivist* approach. According to Achterhuis, an important epistemic model of this turn is to be found in Thomas Kuhn’s constructivist approach in the philosophy of science, which produces the idea of a natural co-evolution between technology and society (see Achterhuis 2001, 6). Manuel Castells, the Spanish sociologist and father of the ‘network society’, gives us a good explanation of the Kuhnian inspiration for the new approach to the question of technology. He affirms that ‘we start from a rejection of technological determinism, as technology cannot be considered independently of its social context’ (Castells 2004, xvii) and that ‘the dilemma of technological determinism is probably a false problem, since technology is society, and society cannot be understood or represented without its technological tools’ (Castells 2010, 5).

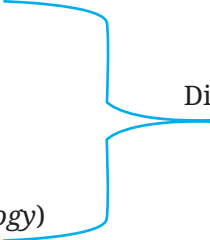
The empirical turn, namely the *second generation of philosophers of technology*, involves scholars such as Albert *Borgmann* (1984), author of the so-called *device paradigm*; Hubert *Dreyfus* (1992), a pioneer of ‘the Critique of Artificial Reason’; Andrew *Feenberg* (1991), who studied with Herbert Marcuse and proposes a *critical constructivism*; Donna *Haraway* (1991), who deals with the question of technology in its link with feminism and posthumanism; the already mentioned Don *Ihde* (1993); and Langdon *Winner* (1980), ‘the

political theorist of technology’ – as defined by Babette Babich (2012–13, 60). These authors are directly considered in Achterhuis’s 2001 book, but various other scholars can be included within the empirical turn, such as Carl *Mitcham* (1994), considered by Achterhuis (2001, 4) ‘the most important historian of the philosophy of technology’; Paul *Durbin*, another significant historian of the philosophy of technology (Durbin and Rapp 1983); Joseph *Pitt* (1995), a point of reference for the engineering-oriented philosophy of technology; David *Noble* (1997), a pioneer of studies about the religious meaning/power acquired by technology as epochal phenomenon; Thomas *Hughes* (1983) and Melvin *Kranzberg* (1985), the founders of the Society for the History of Technology; and Dutch scholar Peter-Paul *Verbeek* (2005), who in the last few years has been the primary follower of Ihde’s postphenomenological approach on the continent. This means that ‘empirical turn’ is no longer a synonym for ‘American philosophy of technology’.

**Second generation**  
(Empirical turn)

- Albert **Borgmann** (*device paradigm*)
- Hubert **Dreyfus** (*Critique of Artificial Reason*)
- Andrew **Feenberg** (*critical constructivism*)
- Donna **Haraway** (*techno-feminism*)
- Don **Ihde** (*postphenomenology*)
- Langdon **Winner** (*political philosophy of technology*)

Directly considered  
by Achterhuis



- 
- Carl **Mitcham** (*‘the most important historian of the philosophy of technology’*)
  - Paul T. **Durbin** (*history of philosophy of technology*)
  - Joseph C. **Pitt** (*philosophy of technology and engineering*)
  - Thomas **Hughes** and Melvin **Kranzberg** (*history of technology*)
  - David **Noble** (*techno-theology*)
  - Peter-Paul **Verbeek** (*continental postphenomenology*)

By commenting on the spirit of the *empirical turn* as expressed by Peter Kroes and Anthonie Meijers (2000), Franssen et al. (2016, 1) affirm that its claim was ‘a reorientation of the community of philosophers of technology toward the practice of technology and, more specifically, the practice of engineering’.

According to Franssen et al. (2016), the very aim of the empirical turn is ‘to steer the philosophical study of technology *away from broad abstract reflections on technology as a general phenomenon* toward addressing philosophi-

cal problems that can be related directly to “*the way technology works*” or to “technology in the making”. In doing so, it focused primarily on the work of engineers’ (Franssen et al. 2016, 2). On this basis, Brey (2010, 40) believes that ‘it is more proper to speak of two empirical turns: in the 1980s and 1990s two distinct approaches have emerged in response to the classical tradition, that both have been claimed to involve an empirical turn’. According to Brey, these two versions of the empirical turn are:

1) a ‘first Empirical Turn’, which can be considered its light version, a ‘*society-oriented* approach in the philosophy of technology’; and

2) the ‘other Empirical Turn’, its hard version. ‘The other empirical turn [...] is instead *engineering-oriented*’ (Brey 2010, 40).

As Brey (2010, 40) explains, the first (society-oriented) empirical turn emerged in the 1980s and 1990s

when more and more philosophers working within the classical tradition were breaking free from some of its assumptions and methods. Neo-Heideggerians, neo-Critical Theorists and post-phenomenologists started to focus on concrete technologies and issues, attempted to develop contextual, less deterministic theories of technology or started borrowing them from STS, and started to assume a less dystopian, more pragmatic and balanced attitude towards modern technology.

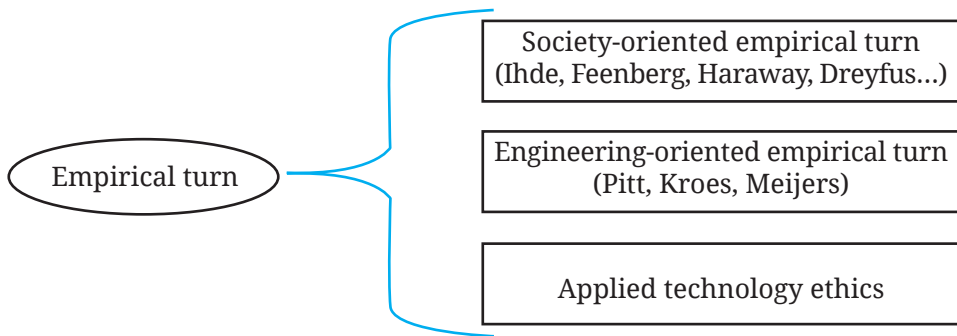
Brey (2010, 40) identifies Andrew Feenberg, Don Ihde, Hubert Dreyfus and Donna Haraway as referent authors of this approach, as well as Larry Hickman, Andrew Light and Bruno Latour. In book form, the manifesto for such an approach is Achterhuis’s *American Philosophy of Technology: The Empirical Turn* (2001).

The very aim of the second (engineering-oriented) empirical turn is ‘to understand and evaluate the practices and products of engineering, rather than anything that happens beyond in society [...] Its primary aim is to understand and evaluate the practices and products of engineering, rather than anything that happens beyond in society.’ This other empirical turn primarily took place in the 1990s and 2000s. It was also ‘borne out of dissatisfaction with the classical approach, but the dissatisfaction was more radical’ (Brey 2010, 40). Brey identifies Joseph Pitt, Peter Kroes and Anthonie Meijers as referent authors of this second approach, while ‘important milestones in this new approach’ (Brey 2010, 41) were *New Directions in the Philosophy of Technology* (Pitt 1995) and *The Empirical Turn in the Philosophy of Technology* (Kroes and Meijers 2000).

Finally, together with these two versions of the empirical turn (‘resulting from an empirical turn within the field’), Brey finds a third approach, that is, the ‘*applied technology ethics*’ that emerges ‘alongside the other two’. In his view, ‘these three approaches now largely define the field’ (Brey 2010, 42).

---

## Brey's version of the empirical turn



### 3. Empirical turn as ontophobic turn

What I am going to argue is that *after 35 years* – taking 1984 as a conventional birth date, namely the year in which Albert Borgmann’s book *Technology and the Character of Contemporary Life: A Philosophical Inquiry*<sup>2</sup> was published – a first attempt at *critical historicisation* of such an experience is possible, and probably useful.

Such a statement needs to be justified, however, as by now there are many other works dealing with a historical evaluation of the empirical turn. Otherwise, as said, the peculiarity – and, hopefully, the usefulness – of my proposal has to do with its critical inspiration, namely its aim to *deconstruct* the empirical turn’s narrative/discourse by calling into question some of its basic and unexpressed assumptions. In my opinion, the other historical overviews of this experience (Brey 2010; Franssen et al. 2016; ...), despite possessing indubitable qualities, are almost always characterised by an acritical attitude which turns them into *historical apologies*, namely confirmations that the empirical turn has been not only a good option (an improvement) for the philosophy of technology but its *only* possible option. In its essence, this kind of historicisation equates to a naturalisation of the empirical turn, which emerges in the end as a matter of fact or even a destiny.

With reference to my critical approach, my hermeneutic hypothesis is that during these 35 years the empirical turn has proven to be an *ontophobic turn*. By this expression I am suggesting an *overreaction* to the so-called essentialist approach to the question of technology, in particular a kind of rejection of Heidegger’s legacy. I will immediately clarify this crucial point of my argument.

---

<sup>2</sup> See Borgmann 1984. This book – and the figure of Borgmann, the ‘German-American philosopher of technology’, generally – represent a natural trait d’union between the continental/Heideggerian tradition and the American philosophical milieu. On Borgmann’s work see Tjmes (2001).

The overreaction against Heidegger's legacy consists in a *two-stage process*. On one side we have the rejection (we could call it a *legitimate rejection*) of the potential 'mystical drift' involved in Heidegger's approach, namely his interpretation of technology as an *Ereignis* (event) within the history of Being. This mystical drift can be considered the *pars construens* (the affirmative side) of Heidegger's philosophy of technology expressed in the idea that 'technology is a way of revealing' (Heidegger 1977, 12). Such a legitimate rejection corresponds to a physiological parricide by the second generation of scholars, in order to free itself from a quite bulky – maybe too bulky – legacy.

However, on the other side this physiological parricide gradually turned into a total refutation: a real *damnatio* which involved the *pars destruens* (the deconstructive side) of Heidegger's approach as well. That is to say, a rejection of what – at least in my opinion – represents the *basic epistemic assumption of the philosophy of technology itself*, namely the condition of possibility for a properly philosophical approach to the question of technology. Such a *pars destruens* is expressed in another well-known Heideggerian sentence, according to which 'the essence of technology is by no means anything technological' (Heidegger 1977, 4). Melvin Kranzberg's first law of technology expresses this point (i.e. the inborn plurivocity/ambiguity of technology) even better by affirming that 'technology is neither good nor bad, nor is it neutral' (Kranzberg 1985, 50).

In my view, this second rejection should be considered an *illegitimate rejection*, that is, an *overreaction* from the second generation of scholars against Heidegger's legacy. Although one can find evidence for this overreactive tendency at every step of the empirical turn, I think it finds its fulfilment – its methodological implementation, so to say – in Ihde's postphenomenological approach.

#### 4. 'Mr Wolfe Syndrome', or engineering as worldview

Concretely, this overreaction can be described as the transition *from an over-distance to an over-proximity*. That is to say, on the one side we have a *disinterest in* – or indifference towards – *the ontic dimension* (namely, the social, political and practical implications) *of technology* and therefore an over-distance. This attitude is typical of the first generation of philosophers of technology and can be epitomised by Heidegger's rejection of 'the instrumental and anthropological definition (*Bestimmung*) of technology' (Heidegger 1977, 5). This disinterest gradually turned into an almost *exclusive interest in the same ontic dimension* (and therefore an over-proximity), a movement typical of the second generation of philosophers of technology. The natural consequence of this attitude is an *a priori disinterest in any ontological implication of technology*, which is characterised *ipso facto* as 'essentialist' or 'deterministic' and thus ends up becoming a taboo. That is to say, a real *onto-phobia*.

The benchmark of this change of attitude in the philosophy of technology is the lexical replacement of its object: the transition *from 'technology'* (in the



---

singular) to ‘technologies’ (in the plural). Not by chance, in his foreword to the English translation of Achterhuis’s book, Don Ihde affirms that precisely this replacement of ‘technology’ with ‘technologies in their relational and contextual implications’ (Ihde 2001, viii) represents a distinguishing feature of the empirical turn.

I agree with the idea that such a replacement means much more than a lexical change, but in my view this semantic surplus corresponds to our *increasing inability to acknowledge technology as something in itself/as such*. In particular, I consider the main outcome of this replacement/inability to be what I call *Mr Wolfe Syndrome*. This formula is inspired by Harvey Keitel’s famous character in Quentin Tarantino’s movie *Pulp Fiction* (1994). This character presents himself as someone who ‘solves problems’. By using this expression, I am therefore referring to the gradual *transformation of the philosophy of technology into a problem-solving activity* or, better, to the fact that philosophers of *technologies* are today becoming (or aspiring to become) ‘guys who solve problems’. Charles Snow, in his famous 1959 report on the two cultures, in an attempt to describe the natural snobbery of the humanities against the sciences (i.e. of humanists/men of letters against scientists/engineers) stated that ‘intellectuals, in particular literary intellectuals, are natural Luddites’ (Snow 2012, 22). Sixty years later, we must admit that those Luddites have turned into strikebreakers.

Mr Wolfe Syndrome embodies the effect of a further and deeper cause, that is, an *epistemic metamorphosis of the philosophy of technology* and, more generally, of philosophy itself. It is the attempt to definitively make it a ‘*positive Wissenschaft*’ (*positive science*), that is, a knowledge grounded on a ‘*positum*’: an absolute givenness, an epistemic and ontological dogma which can no longer be questioned. I mean ‘*positum*’, ‘positive character’ (*Positivität*) and ‘positive science’ according to their interpretation in Heidegger’s essay *Phenomenology and Theology* ([1927] 1998): his epistemic manifesto. Here (Heidegger 1998, 41), he gives the following epistemic-ontological definition of the positive sciences.

there are two basic possibilities of science: sciences of beings, of whatever is, or ontic sciences, and *the science of being, the ontological science, philosophy*. Ontic sciences in each case thematize a given being that in a certain manner is always already disclosed prior to scientific disclosure. We call the sciences of beings as given – of a *positum* – positive sciences.

On the contrary, he continues, ‘ontology or the science of being [...] demands a fundamental shift of view: from beings to being’. As a consequence, ‘within the circle of actual or possible science of beings – the positive sciences – there is between any two only a relative difference [...] On the other hand, every positive science is *absolutely*, not relatively, different from philosophy’ (Heidegger 1998, 41). Given these assumptions, by Mr Wolfe Syndrome I mean the attempt to definitively make disappear, within the framework of the topic ‘technology’, the epistemic biodiversity of philosophy; to make it unrecognis-



able, unperceivable. Or, if you prefer, to make us once and for all blind to this kind of difference.

In turn, both this epistemic metamorphosis and the consequent Mr Wolfe Syndrome can be considered the final results produced by the *eclipse of the epistemic difference between 'problem' and 'question'*. By 'problem' I mean that kind of interrogation which allows only *solution* as its possible answer. And by 'solution' I mean that kind of answer which completely annihilates its own interrogation. That is to say, after reaching its solution, the interrogation in itself disappears, becomes nothing, ceases to make sense precisely because it is entirely solved. Problem is nothing but the premise (i.e. the occasion, the pretext) of a solution. On the other side, with 'question' (or better 'basic question' – I refer here to the German word *Frage*, or better *Grund-Frage*) I mean a kind of interrogation whose answer can be something different from a solution. A question is a potential unsolvable interrogation. Possible examples of these questions as unsolvable interrogations are two philosophical *Grundfragen par excellence*, that is, 'Why is there something, rather than nothing?' and 'What is called thinking?'. In the latter case, an 'adequate' answer (namely a pathic, non-logic, pre-logic answer) could be that philosophical keyword which Plato (*Theaetetus*, 155d) and Aristotle (*Metaphysics* I, 2, 982b) already suggested as the origin of thought: '*thaumazein*'. *Thaumazein* represents a paradigmatic example of an answer without solution, that is, an answer which keeps its own interrogation alive, leaves it open. As a result, 'question' equates to an unsolvable but not meaningless interrogation.

On this basis, I believe that technology as philosophical issue (namely, as historical/epochal phenomenon) equates to such a *Grundfrage*. My worry is that, after 35 years, the empirical turn as ontophobic turn (i.e. overreaction against Heidegger's legacy) could entirely eclipse the epistemic difference between question and problem, and thus make any *Grundfrage* impossible. That is to say, it could make us insensitive, blind to any *Grundfrage*. In other words, if we firmly believe that any question must require/imply a solution (namely, that any question exists only insofar as it implies a solution, that any question must become a problem), then an unsolvable question (that is, a *Grundfrage*) becomes a non-question, namely a pseudo-problem, a pure mistake or misunderstanding. Translated as an ontological formula, this approach would read: '*What cannot be solved, is not.*' As an imperative, it would read: '*Make everything solvable.*'; '*Make a problem of everything.*' *Solveability* therefore emerges as an epochal *paspartout*, the current basic ontological feature of any entity.

Now, insofar as the problem-solving logic represents the conceptual dispositive of the engineering approach, I call such an attempt to eclipse the epistemic difference between problem and question *engineerisation*. The ultimate goal of this engineerisation is to achieve a complete *problematization of reality*, that is, to build an epistemic and ontological framework within which 'problem' becomes the only possible way of interrogation. On this ba-

---

sis, the ‘question concerning technology’, as *Grundfrage*, is bound to become a non-question, a nonsense. It is not by chance therefore that Brey and the other apologetic historians of the empirical turn identify its peculiarity precisely with the definitive approximation of the philosophy of technology to engineering. In particular, according to Brey (2010, 40), the engineering-oriented empirical turn represents the authentic empirical turn, its natural and necessary outcome. That is to say, its entelechy.

## 5. A requiem for the *philosophy of technology*?

With reference to this whole state of things I have described, *my objection* is the following. If Technology (with a capital ‘T’) as such, that is, technology as potential *Weltanschauung* or *grand récit* of our age, as current ‘subject of history’ (Anders 1992, 271–9) ... well, if *technology as such is/becomes nothing* (if it comes to represent at most the umbrella term or the summation of the single technologies), then the paradoxical but entirely consequential result of this situation is that the philosophy of technology ceases to have a meaning and a value in itself. In other words, if the philosophy of technology turns entirely into a problem-solving activity (into a search for solutions in front of the concrete problems emerging from the single technologies), then it must be admitted that this kind of activity can be performed much better by ‘experts’ (scientists, engineers, politicians ...) than by philosophers.

As a consequence, the ontophobic turn in philosophy of technology (namely, the overreaction against Heidegger’s legacy) culminates in the disappearance of the reason itself for a strictly philosophical approach to the question of technology. Given this assumption, the paradoxical *accomplishment/fulfilment of the empirical turn* should be the final *self-suppression*, or at least *self-overcoming*, of the *philosophy of technology*.

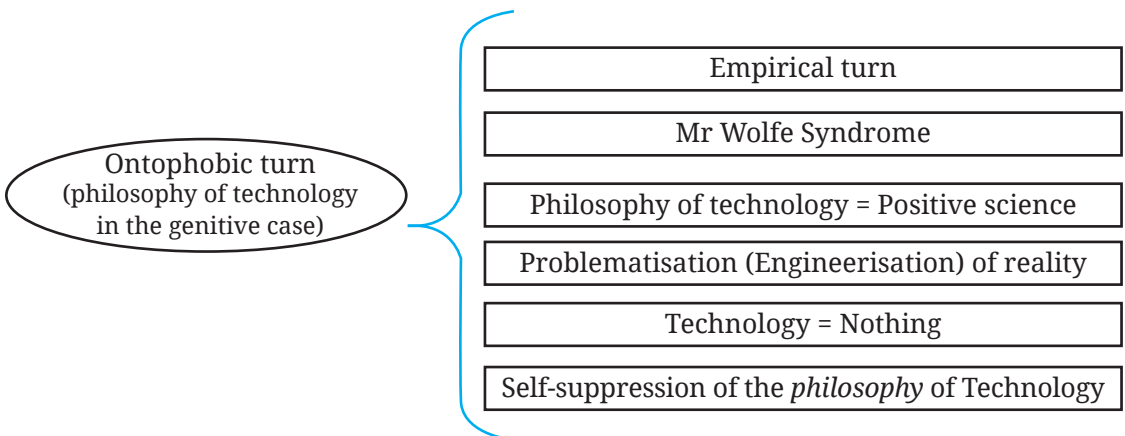
This objection also gives me the opportunity to emphasise a significant aporia within Brey’s argument and more generally within the discourse of the apologetic historians of the empirical turn. Dealing with the ‘Limitations of Contemporary Philosophy of Technology’ (the premise for establishing an ‘Agenda for the Philosophy of Technology’), Brey focuses on three questions which he considers the ‘major questions’ for the philosophy of technology. They are: 1) ‘*What is technology?*’; 2) ‘*How can the consequences of technology for society and the human condition be understood and evaluated?*’; and 3) ‘*How should we act in relation to technology?*’ (Brey 2010, 43 – my italics). The first question is ‘the central concern of engineering-oriented philosophy of technology’, the second question is ‘the province of society-oriented philosophy of technology, but also of technology ethics’, while the third is ‘wholly the concern of technology ethics’ (Brey 2010, 43). Brey affirms that only the engineering-oriented philosophy of technology is able to carry out its task, that is, to answer (solve) its own question (problem), while in their current versions the society-oriented philosophy of technology

and the technology ethics ‘are not sufficiently equipped to provide full and cogent answers to the second and third research question[s]’ (Brey 2010, 43).

In my view, the basic aporia of this position is that Brey presumes to answer a question/solve a problem – ‘What is technology?’ – which he himself (through his approach to the question of technology) has made meaningless. More clearly, my point is: how can answering the question concerning the ontological status, or even the essence, of technology (in the singular, as something in itself) be the same approach which characterises itself by establishing the definitive overcoming of ‘Technology-with-a-capital-T’? That is to say, by establishing the definitive overcoming of the ontological question itself? The only possible *escamotage* I can find for this aporia has to do with the formulation of the question, that is, the interpretation of its meaning. The engineering-oriented approach – which in turn represents the quintessence of the empirical turn – can answer the question ‘What is technology?’ only because, according to its assumptions, ‘*What is?*’ *ipso facto* means ‘*How does it work?*’. As a result, the ontological question ‘What is technology?’ turns into the concrete problem ‘How does (a single) technology work?’ and thus immediately becomes something solvable, that is, the only legitimate/real interrogation for this approach. It is a paradigmatic example of empirical turn as ontophobic turn, because this reformulation/translation of the question ‘What is technology?’ is entirely based on a negation/annihilation of the ontological level.

This ontophobic *escamotage* is the attempt to annihilate the epistemic peculiarity of the philosophy of technology (and of philosophy in general) by definitively transforming it into a *positive Wissenschaft* or, better, a problem-solving activity; that is to say, by identifying – (con)fusing – philosophy with engineering. Brey’s aporia, particularly in its engineering-oriented version, unintentionally confirms that the empirical turn’s only possible answer to the question ‘What is technology?’ is ‘Technology (i.e. in itself/as such) is nothing.’.

### Ontophobic turn



---

## 6. The ontophilic turn: Towards a philosophy of technology in the nominative case

Before concluding, I would like to quote some lines from *Franco Volpi* that represent a perfect synthesis of my point, an important source of inspiration for the *pars construens* of my work on the philosophy of technology. In his book on nihilism, Volpi speaks about the risk of *genetivisation* for *philosophy* today, in particular for the philosophy of technology. He affirms:

There is a risk: that yet another philosophy *in the genitive case* will be produced. I mean, a reflection whose only function is ancillary and subordinate [... T]he risk of numerous genitive philosophies is to reduce philosophical thought to a noble *anabasis*, namely to a strategic withdrawn from the great questions to take refuge in problems of detail [...] So, one asks oneself: is *philosophy of technology in the nominative case* (*filosofia della tecnica al nominativo*) possible? (Volpi 2004, 146–7).

On the basis of the arguments I sketched in this paper, I think that such a genetivisation is already underway, and that it corresponds to the ontophobic outcome characterising the current mainstream in the philosophy of technology, that is, the attempt to overcome/annihilate its epistemic peculiarity by transforming it into a *positive Wissenschaft*, or problem-solving activity. Given this assumption, in my view the most urgent work needed in this field is an attempt to give an *affirmative reply to Volpi's question* (a philosophy of technology in the nominative case is possible)<sup>3</sup> by means of a *countermovement* (in the sense of Nietzsche) towards the currently prevailing ontophobic turn.

The first step of an *ontophilic turn* consists in the right metabolisation of Heidegger's legacy. In other words, the foundation of a philosophy of technology in the nominative case must involve a *Heidegger renaissance* or a 'going back to being fair with Heidegger'. This means that we must avoid the potential mystical drift of his approach without compromising the epistemic *imprimatur* he gave to this area of study. What is truly at stake in this *Heidegger renaissance* is both the safeguarding of the *Fragwürdigkeit* (questionworthiness) of *technology for philosophical thought* and the *epistemic peculiarity/biodiversity of philosophy itself*, since a philosophy of *technology* answers not

---

<sup>3</sup> For several years I have been working on my personal interpretation of such a philosophy of technology. I call it 'Philosophy of Technology in the Nominative Case (TECNOM)' and have presented it in various papers, for example Cera 2017 and 2018, 131–80. I would like to mention two of the many heterodox examples of such a countermovement against the current mainstream in the philosophy of technology: 1) Babette Babich who, as Ihde's student, criticises postphenomenology's overreaction against the classical philosophy of technology; and 2) the 'Wageningen-Nijmegen Group' (Vincent Blok, Pieter Lemmens and Jochem Zwier) which claims a 'terrestrial turn in philosophy of technology' (see Lemmens, Blok and Zwier 2017).

only the question ‘What is technology?’ but also – and maybe even more so – the question ‘What is *philosophy*?’. At the basis of any philosophical interpretation of technology lies an interpretation of philosophy.

If technology as such/in itself is something, in particular if it – as epochal phenomenon – equates to the current subject of history, then the philosophy of technology will also emerge as the current version of the philosophy of history. Or, better, it will emerge as our best resource for doing in the here and now what philosophy has always tried to do: to ‘comprehend its own time in thoughts’ (Hegel 1991, 21).

At this point it should be clear that the countermovement I am proposing (i.e. ontophilic turn or philosophy of technology in the nominative case) consists of a *re-philosophising of the philosophy of technology*; that is, a *philosophical (re)turn in the philosophy of technology*.

As a conclusion, I would like to quote the wise words that Albert Borgmann shared with me during a private conversation on these topics (as is well-known, Borgmann is one of the protagonists of the empirical turn, a key figure in the transition from the first to the second generation of philosophers of technology). I think they represent the real spirit of my proposal, which wants to be not an ordeal – that is, a fanatic *pro* or *contra* Heidegger – but an attempt to preserve the *irreplaceability of a strictly philosophical approach to the question of technology*. My proposal’s very aim is ‘only’ to keep the *difference between question (Grundfrage) and problem* alive, to keep our sensibility towards such a *nuance* alive. In fact, ‘by honoring this questionworthiness (*Fragwürdigkeit*), philosophy possesses its own dignity, one that cannot be derived from elsewhere and cannot be calculated’ (Heidegger 2012, 7).

Borgmann affirms:

I agree that the *Grundfrage* is the source of the deepest insights and that we should not let it get buried by a problem-oriented approach. There are Sundays in philosophy, when we festively celebrate insight. But there is also the week-day philosophy, when we busy ourselves with problems. As long as problem-solving does not obliterate the *Grundfrage*, we should allow for a space for it.

## References

- Achterhuis, Hans, (ed.). *De maat van de techniek: Zes filosofen over techniek, Günther Anders, Jacques Ellul, Arnold Gehlen, Martin Heidegger, Hans Jonas en Lewis Mumford*. Ambo: Baarn & Schoten, 1992.
- Achterhuis, Hans (ed.). *Van stoommachine tot cyborg; denken over techniek in de nieuwe wereld*. Amsterdam: Uitdigeverij Ambo, 1997.



- 
- Achterhuis, Hans (ed.). *American Philosophy of Technology: the Empirical Turn* (The Indiana Series in the Philosophy of Technology), trans. R. P. Crease. Bloomington/Indianapolis: Indiana University Press, 2001.
- Anders, Günther. *Die Antiquiertheit des Menschen 2. Über die Zerstörung des Lebens im Zeitalter der dritten industriellen Revolution*. München: Beck, 1992.
- Babich, Babette. "O, Superman! Or Being towards Transhumanism: Martin Heidegger, Günther Anders, and Media Aesthetics." *Divinatio* XXXVI (2012-2013): 41–99.
- Borgmann, Albert. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Chicago: University of Chicago Press, 1984.
- Brey, Philip. "Philosophy of Technology after the Empirical Turn." *Techné: Research in Philosophy and Technology* 14, no. 1 (2010): 36–48. <https://doi.org/10.5840/techne20101416>.
- Castells, Manuel (ed.). *The Network Society: A Cross-cultural Perspective*. Cheltenham (UK), Northampton (MA): Edward Elgar, 2004.
- Castells, Manuel. *The Rise of the Network Society* (The Information Age: Economy, Society and Culture Vol. I). Malden (MA), Oxford (UK): Wiley-Blackwell, 2010.
- Cera, Agostino. "The Technocene or Technology as (Neo)environment." *Techné: Research in Philosophy and Technology* 21, no. 2/3 (2017): 243–81. <https://doi.org/10.5840/techne201710472>.
- Cera, Agostino. *Der Mensch zwischen kosmologischer Differenz und Neo-Umweltlichkeit. Über die Möglichkeit einer philosophischen Anthropologie heute*. Nordhausen: Verlag Traugott Bautz, 2018.
- Dreyfus, Hubert. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge (Mass.): MIT Press, 1992.
- Durbin, Paul and Friederich Rapp (eds.). *Philosophy and Technology* (Boston Studies in the Philosophy of Science vol. 80). Dordrecht and Boston: D. Reidel Publishing Company, 1983.
- Ellul, Jacques. *The Technological Society*, trans. J. Wilkinson, New York: Vintage Books, 1964.
- Feenberg, Andrew. *Critical Theory of Technology*. Oxford: Oxford University Press, 1991.
- Feenberg, Andrew. *Questioning Technology*. London/New York: Routledge, 1999.
- Franssen, Maarten and Pieter E. Vermaas and Peter Kroes and Anthonie W.M. Meijers (eds.). *Philosophy of Technology after the Empirical Turn* (Philosophy of Engineering and Technology vol. 23). Cham: Springer International, 2016.
- Gehlen, Arnold. *Man in the Age of Technology*, trans. C. McMillan and K. Pillemer. New York: Columbia University Press, 1980.
- Haraway, Donna. *Simians, Cyborgs and Women: The Reinvention of Nature*. New York: Routledge, 1991.
- Hegel, Georg Wilhelm Friedrich. *Elements of the Philosophy of Right*, trans. H. B. Nisbet. Cambridge and New York: Cambridge University Press, 1991.
- Heidegger, Martin. "The Question Concerning Technology." In *The Question Concerning Technology and Other Essays*, trans. W. Lovitt, 3–35. New York & London: Garland, 1977.
- Heidegger, Martin. "Phenomenology and Theology." In *Pathmarks*, trans. W. McNeill, 39–62. Cambridge & New York: Cambridge University Press, 1998.
- Heidegger, Martin. *Contribution to Philosophy (Of the Event)*, trans. R. Rojcewicz and D. Vallega-Neu. Bloomington/Indianapolis: Indiana University Press, 2012.
- Hughes, Thomas P. *Networks of Power: Electrification in Western Society, 1880-1930*. Baltimore: Johns Hopkins University Press, 1983.



- Ihde, Don. *Postphenomenology: Essays in the Postmodern Context*. Evanston (Ill.): Northwestern University Press, 1993.
- Ihde, Don. "Foreword." In *Achterhuis 2001*: vii–ix.
- Kranzberg, Melvin. "The information Age: Evolution or Revolution?" In *Information Technologies and Social Transformation*, edited by Bruce R. Guile, 35–53. Washington (DC): National Academy Press, 1985.
- Kroes, Peter and Anthonie Meijers (eds.). *The Empirical Turn in the Philosophy of Technology*. Amsterdam: JAI-Elsevier, 2000.
- Lemmens, Pieter and Vincent Blok and Jochem Zwier. "Toward a Terrestrial Turn in Philosophy of Technology." *Techné: Research in Philosophy and Technology* 21, no. 2/3 (2017): 114–26. <https://doi.org/10.5840/techne2017212/363>.
- Mitcham, Carl. *Thinking through technology: The path between engineering and philosophy*. Chicago: University of Chicago Press, 1994.
- Noble, David. F. *The Religion of Technology; The Divinity of Man and the Spirit of Invention*. New York: Knopf, 1997.
- Pitt, Joseph T. (Ed.). *New Directions in the Philosophy of Technology (Philosophy and Technology vol. 11)*. Dordrecht: Springer, 1995.
- Rosenberger, Robert and Peter-Paul Verbeek (eds.). *Postphenomenological Investigations: Essays on Human-Technology Relations*. Lanham (MD): Lexington Books, 2015.
- Selinger, Evan (Ed.). *Postphenomenology: A Critical Companion to Ihde*. Albany: State University of New York Press, 2006.
- Snow, Charles P. *The Two Cultures*. Cambridge and New York: Cambridge University Press, 2012.
- Tijmes, Pieter. "Albert Borgmann: Technology and the Character of Everyday Life." In *Achterhuis 2001*: 11–36.
- Verbeek, Peter-Paul. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State: Penn State University Press, 2005.
- Volpi, Franco. *Il nichilismo*. Roma-Bari: Laterza, 2004.
- Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–36.